Variants of compound models and their application to citation analysis

Wan Jing Low

A thesis submitted in partial fulfilment of the
requirements of the University of Wolverhampton
for the degree of Doctor of Philosophy

March, 2017

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

Signature: .....
Date: .....

Abstract

This thesis develops two variant statistical models for count data based upon compound models for contexts when the counts may be viewed as derived from two generations, which may or may not be independent. Unlike standard compound models, the variants model the sum of both generations. We consider cases where both generations are negative binomial or one is Poisson and the other is negative binomial. The first variant, denoted SVA, follows a zero restriction, where a zero in the first generation will automatically be followed by a zero in the second generation. The second variant, denoted SVB, is a convolution model that does not possess this zero restriction. The main properties of the SVA and SVB models are outlined and compared with standard compound models. The results show that the SVA distributions are similar to standard compound distributions for some fixed parameters. Comparisons of SVA, Poisson hurdle, negative binomial hurdle and their zero-inflated counterpart using simulated SVA data indicate that different models can give similar results, as the generating models are not always selected as the best fitting.

This thesis focuses on the use of the variant models to model citation counts. We show that the SVA models are more suitable for modelling citation data than other previously used models such as the negative binomial model. Moreover, the application of SVA and SVB models may be used to describe the citation process.

This thesis also explores model selection techniques based on log-likelihood methods, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The suitability of the models is also assessed using two diagrammatic methods, randomised quantile residual plots and Christmas tree plots. The Christmas tree plots clearly illustrate whether the observed data are within fluctuation bounds under the fitted model, but the randomised quantile residual plots utilise the cumulative distribution, and hence are insensitive to individual data values. Both plots show the presence of citation counts that are larger than expected under the fitted model in the data sets.

This thesis is dedicated to my parents.

Acknowledgement

Firstly, I would like to thank my amazing supervisory team, Dr. Paul Wilson (director of studies) and Professor Mike Thelwall, for their guidance, kindness, patience, inspiration, generosity, encouragement and invaluable advice throughout my studies. They have been great role models. I am very lucky to have two supervisors who are extremely knowledgeable and supportive. Their expertise and enthusiasm for research have been a great motivation for me. I would also like to express my gratitude to Professor Mike Thelwall for providing the research funding to sponsor my PhD and supported the attendance to conferences which further facilitated my professional development.

I would also like to thank Dr. Liam Naughton for proof reading my thesis and Dr. David Huen for allowing me to use his Linux system so that I could work more efficiently. I would also like to thank the members of the Statistical Cybermetrics Research Group for their feedback and the interesting discussions regarding my research during the monthly forums. I would also like to thank my friends for making this journey more pleasant.

Finally, I would also like to extend my deepest gratitude and appreciation to my wonderful parents and family, especially my sister, for their endless love, continuous support and encouragement throughout my studies. It would not be possible for me to complete this work without them.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

In recent years, the increasing volume of count data from economics and science has emphasised the importance of statistical distributions to model counts. Count data concern the number of events occurring for a fixed period of time. Poisson and negative binomial count distributions are commonly used to model this type of data. The Poisson model assumes the equality of mean and variance, but violations of this assumption are common in real data. The negative binomial and Poisson Inverse Gaussian (PIG) models have been used when the variance is greater than the mean (Puig and Valero, 2006). It is also possible for the variance to be less than the mean but this is rare (Cox, 1983).

This thesis focuses on citation counts, which are the number of citations received by academic publications, such as journal articles. Citation counts have been widely used as an indicator for the impact of research publications by those funding, managing or conducting research. It is therefore important to identify the most appropriate statistical model for citation data to maximise the power and validity of future analyses, and to better understand the citation process (Price, 1976).

The next section in this chapter discusses the use of citation counts for research evaluation, including a discussion of bibliographic databases and previously used statistical models in citation analysis. Section 1.2 presents the research gaps, aims and research questions. Finally in Section 1.3, the structure of this thesis is outlined.

1.1 Citation counts for research evaluation

Citation counts are used to support peer review judgements in some subject areas in the Research Excellence Framework (REF), which replaced the Research Assessment Exercise (RAE) after 2008 (HEFCE, 2009). The REF is used by the Higher Education Funding Council for England (HEFCE) in the United Kingdom

to measure the quality of research in universities (Smith et al., 2011). Hence, fitting a suitable statistical model will potentially maximise the predictive power of citation counts. Citation counts also play a major role in university rankings (Liu, 2009; Rauhvargers, 2011). These rankings affect student and researcher recruitment to institutions, which is vital for their continued prosperity. In addition, some institutions use citation counts whilst formulating research policies (Vieira and Gomes, 2010) and to evaluate individuals. An example is the h-index (or Hirsch-index) (Hirsch, 2005; Radicchi and Castellano, 2013). A researcher has a h-index of x, if x of the researcher's papers each has at least x citations, where x is the highest possible such integer. Citation counts have also been used to evaluate departmental performance across different universities (Kinney, 2007). On an individual level, citation counts may be used to evaluate the performance of researchers for hiring or promotion purposes (Vieira and Gomes, 2010). Citation counts may also affect decisions about grant applications (Bornmann and Daniel, 2006). Although the use of citation counts to support research evaluation is controversial and also inappropriate in some disciplines (for example, arts, humanities, perhaps also formal sciences), they are a widely used fact of life in academia.

Citation patterns vary between disciplines; for example social science and humanities fields tend to have lower citation counts than medicine (Rauhvargers, 2011). There are also vast differences in citation behaviours within science subjects. For example, biology papers are typically more cited than physics papers (Kinney, 2007). Hence, citation counts are only comparable within a field, but not across fields. It would be advantageous to study and fit appropriate statistical distributions to as many fields as possible to further enhance the general understanding of the citation process.

1.1.1 Bibliographic databases

Citation data often consist of a list of publications with associated meta data, such as authors, year, journal name and title. These can be obtained from bibliographic databases, which mainly hold collections of published literature such as journal articles, review articles, conference proceedings, patents and books (Feather and Sturges, 2003). Examples of bibliographic databases are the Web of Science (previously known as the Web of Knowledge), Scopus and Google Scholar, where the latter two are alternatives indexes introduced in 2004 (Leydesdorff and Milojević, 2012). Unlike Web of Science and Scopus, Google Scholar does not have a download feature that allows users to download a set of entries containing publication details.

Scopus has wider coverage than the Web of Science, especially for referenced documents (Vieira and Gomes, 2009). Although Google Scholar has the most comprehensive coverage amongst the three databases, it contains large numbers of duplicate citations (Harzing and Alakangas, 2016). Google Scholar may also include blogs or magazine articles because it lacks quality control (Harzing and Alakangas, 2016). Therefore, the citation data in this thesis were obtained from Scopus, based on specific subject areas and publication years.

1.1.2 Statistical models used for citation analysis

Various statistical models have been used for citation analysis, including Lotka's law for scientific productivity (Lotka, 1926) and the power law (Price, 1965). Before fitting these, articles with small numbers of citations are often excluded. This step is most frequently used to conform to the assumptions used by the power law model. Recently, negative binomial models have also been used for citation analysis (e.g. Bornmann and Daniel, 2008; Sud and Thelwall, 2016) as citation data are often left skewed, with variance bigger than the mean (Bornmann and Daniel, 2016), relative to a Poisson model. Other models such as the zero-inflated and hurdle models have also been considered (Lee et al., 2007; Didegah and Thelwall, 2013b). These models will be discussed in detail in Chapter 2. A major advantage of these models is that publications with few citations do not need to be excluded and so the models can be more realistic and have much greater practical value.

1.2 Aims and research questions

This thesis introduces and tests new statistical models for count data, which are variants of compound (also known as stopped sum) models. The motivation behind this is that not enough attention has been paid to the selection and validation of appropriate statistical methodologies for citation counts. Moreover, measurements in information science are often ambiguous (Bookstein, 2001), highlighting the challenges faced by any statistical analysis.

Although various statistical models have been fitted to citation count data, such as those mentioned in Section 1.1.2, there is no consensus on which is "best" and more studies are needed. This thesis aims to partly fill this gap by investigating several new models.

Moreover, there has been a lack of emphasis on statistical methods when fitting citation count data. This thesis aims to fill this second gap by introducing a novel method for testing the fit of citation count data which also has a practical

interpretation. Thus, this thesis focuses on model development and fitting. In particular, it assesses whether compound models, such as the Neyman type A, and variants of compound models are suitable for citation data sets. This includes developing R code to fit the proposed variant models and investigating their functionality.

Although model selection techniques such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) provide good evidence, it may be unwise to rely solely on single goodness of fit values because some models may be too flexible (Einbeck and Wilson, 2016). Hence, apart from standard model selection criteria, this thesis also includes the use of randomised quantile residual plots (Dunn and Smyth, 1996) and Christmas tree plots (Einbeck and Wilson, 2016) to assess model fits. In summary, this thesis aims to:

(i) Assess a range of statistical distributions for modelling citation data.

(ii) Introduce variants of compound models and deduce their main properties.

(iii) Investigate the appropriateness of these variant models in modelling count data with an emphasis on citation data.

(iv) Assess the effectiveness of diagnostic plots as a model validation technique for citation count data.

Thus, our research questions are:

(i) Are compound models appropriate for modelling citation data?

(ii) Are the proposed variants of compound models suitable for citation analysis?

(iii) Are randomised quantile residual plots and Christmas tree plots more useful than AIC and BIC for model validation in citation analysis?

(iv) Can the proposed variants of compound models be extended to incorporate covariates?

1.3 Thesis structure

This thesis consist of 7 chapters.

In Chapter 2, Preliminary concepts, some count models used in scientometric and standard compound distributions are discussed in detail. Some common model selection and validation techniques are also reviewed.

Chapter 3, Variants of compound models, introduces the SVA and SVB variants of compound models. The properties of these variants, such as their moment generating functions, probability generating functions, characteristic functions, expectations, variances, skewness and kurtosis coefficients are discussed. The code used to calculate the probabilities of the standard compound models and the proposed variant models is also described. Given that the model fitting process involves estimations of parameters, optimisation methods used when fitting these models are also described.

Chapter 4, Simulation studies, is split into three main parts. The first part describes simulation studies using standard compound distributions while the second uses their variants. The third part compares one of the variant distribution, SVA, with the hurdle and zero-inflated models.

Chapter 5, Applications, consists of three applications of the variant models. First, the variant models are used to model citation counts without incorporating any covariates. The second part extends the initial citation analysis by incorporating two covariates. The variant models are then used in biodosimetry analysis and their fits compared to the standard compound models.

Chapter 6, Christmas tree plots for model validation, assesses an alternative model checking technique. The plots are first applied to simulated data fitted to their generating model. The Christmas tree plots are then used to illustrate the suitability of some models used in the analyses in Chapter 5.

Chapter 7, Conclusion, concludes by discussing the novel contributions made in this research, some limitations of this project and proposes future work.

Chapter 2

Preliminary concepts

2.1 Introduction

In this chapter, we provide some background knowledge on the concepts used in this thesis. In Section 2.2, we review some count distributions that have been used in scientometrics. In Section 2.3, we describe some standard compound models that will be used in analyses in the subsequent chapters. Given a list of models, it is important to select a suitable model for future prediction. Hence, in Section 2.4, we discuss some model selection techniques, which include maximum likelihood estimation and two commonly used model selection criteria, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). In addition, we discuss other validation techniques used such as standard errors of parameter estimates and the use of randomised quantile residuals.

2.2 Count models used in scientometrics

Scientometrics is the study of science from a quantitative perspective (Leydesdorff and Milojević, 2012). The most commonly used data type in this field is citation counts, which is our focus in this thesis. The use of statistical models in scientometrics dates back to the 1920s, with Lotka's law on scientific productivity (Lotka, 1926). Since then, various models have been proposed. However, the skewed nature of citation count data (Price, 1951, 1976), with a heavy right tail, adds difficulty to the task of identifying and fitting appropriate statistical models to citation counts (Clauset et al., 2009). In this section, we review some distributions that are frequently used to model citation counts and discuss some other recently used models.

2.2.1 Lotka's law for scientific productivity

Using an index containing Chemistry abstracts for the years 1907 to 1916 and a historical index for Physics, which consists of the names of all important physics contributions up to and including year 1900, Alfred J. Lotka deduced that the number of people writing k scientific papers is proportional to $1/k^2$. Lotka had originally assumed that the proportion of people who make k contribution(s) is $f(k)$, where:

$$f(k;\alpha) = \frac{C}{k^\alpha} \tag{2.1}$$

Here C is a constant, and α varies with the data (Kretschmer and Rousseau, 2001) but Lotka found that α was usually 2. Equation (2.1) is also the general form of a power law (see Section 2.2.2), where the probability of obtaining k is inversely proportional to the power of k. For the special case of $\alpha = 2$, Lotka found that $C = 6/\pi^2$ (which is approximately 0.6). Hence, Lotka concluded that about 60% of all contributors have made exactly one contribution. In other words, 60% of all authors had written just one paper.

Bookstein (1990) concluded that Lotka's law is generally applicable to scientific productivity as this law will hold irrespective of the method used to assign credits to authors when multiple authors contribute to one publication. However, Rousseau (1992) gave a counterexample by studying authors cited in the bibliography of a review paper and a book. Rousseau showed in his example that the law does not hold when using adjusted counts, that is when weights are equally distributed among authors. For example, if a paper has 5 authors, then authors of that paper will each have weights of 1/5, so they each are assumed to have written 1/5 of a paper. Using the total count procedure, with each author receiving full credit, Kretschmer and Rousseau (2001) concluded that the fit of Lotka's Law is only reasonable when articles have up to 40 co-authors.

2.2.2 Power laws

Whilst Lotka's law is commonly used for author productivity distributions, a more general power law has been used to investigate citation counts for sets of publications. However, the power law is only suitable for modelling the tail of citation distributions (Redner, 1998). The power law exists as both continuous and discrete distributions (Clauset et al., 2009). In the continuous case, the observed value, x, is a real number. In the discrete case, x is an integer value. Note that the probability function for a continuous distribution is known as probability density function (pdf) and for a discrete function it is known as probability mass

function (pmf). The pdf and pmf of the power law is similar to Equation 2.1:

$$f(x;\alpha) = Cx^{-\alpha} \qquad \text{for } x \geq x_{min} \tag{2.2}$$

where α is a parameter (also known as the power law exponent) that depends on the data and there must exist a lower threshold, x_{min} for it to be valid. Since we are interested in citation counts, which are discrete, we only consider the discrete power laws where x_{min} is also an integer value. For example, for the top $1,200$ most cited papers published in 1991 in journals that are catalogued in the Institute for Scientific Information catalogue, Redner (1998) considered a citation time frame from 1981 to 1997 and concluded that $x_{min} = 85$, and $\alpha \approx 3$. However, this is based on visual inspections of the observed plots. Thus, if $N(x)$ is the number of papers which have been cited x times, then by plotting $\log N(x)$ against $\log x$, Redner was able to visually estimate the lower threshold value, x_{min}. In addition, α was estimated based on a Zipf plot, that is a plot of the number of citations of the k^{th} ranked paper against rank k on a double logarithmic scale. Hence, these observations are purely based on the author's perception, which may be unreliable. Similarly, Lehmann et al. (2003), observed $x_{min} = 50$, and $\alpha \approx 2.3$ when analysing $281,717$ journal papers in the field of high-energy physics, extracted from the SPIRES data base. Glanzel (2007) stressed that the values of x_{min} and α differ across fields. Albarrán and Ruiz-Castillo (2011) found that although the power law typically models a small percentage of articles, these articles generally receive a large percentage of all citations within a field. A comprehensive guideline to estimate α and fit the power law using maximum likelihood methods was given by Clauset et al. (2009). Nonetheless, given the limitation that a lower threshold is required, it is doubtful that this distribution can fully describe the whole citation range in many fields (Sangwal, 2013; Perc, 2010; Redner, 1998).

2.2.3 The Yule-Simon distribution

The Yule-Simon distribution:

$$f(x;g,s) = \frac{g}{s} \cdot \frac{\Gamma(x) \cdot \Gamma(\frac{g}{s}+1)}{\Gamma(x+\frac{g}{s}+1)} \tag{2.3}$$

was initially derived by Yule (1925) to describe species distribution in a family of organisms, where g is the generic mutation rate (i.e. the rate of production of new species of the same genus) and s is the specific mutation rate (i.e. the rate of production of a new genus) (Garcia, 2011; Simon, 1955). Newman (2005) stated that the tail of the Yule-Simon distribution follows a discrete power law.

Brzezinski (2015) compared the Yule distribution with other discrete distributions, such as exponential, Weibull, Tsallis, digamma, lognormal and the power law with exponential cutoff, and found that the latter two and Yule tends to fit citation count data better than the other fitted models in terms of log-likelihood values. However, similar to the power law, this model only fits above a threshold. It is also common that if the Yule-Simon model, power law or hooked power law (see Section 2.2.4) fit a data set well, then this suggests, but does not prove, that the data may be generated by a preferential attachment process.

2.2.4 The hooked power law

The hooked power law is an extension of the power law, where the variable of interest, say x, is inversely proportional to a power of $(x + B)$, where B is a constant:

$$f(x;\alpha) = \frac{A}{(B+x)^{\alpha}} \tag{2.4}$$

This was originally proposed by Pennock et al. (2002) for web links, but its application has also been extended to citation counts (Eom and Fortunato, 2011; Thelwall and Wilson, 2014a). Unlike the power law, the hooked power law does not require a lower bound for citation count data, as deviations from a pure power law for low values of x can be accomodated by B. The hooked power law is equivalent to the power law if $B = 0$. Thelwall (2016b) stated that the hooked power law is a discrete version of the Lomax distribution (Lomax, 1954), or a special case of the Pareto type II distribution, which is also known as the Pearson type VI distribution (Johnson et al., 1994, p.575).

2.2.5 The lognormal distribution

Rousseau (1992) suggested that the lognormal distribution could fit citation data reasonably well. However, some adjustments need to be applied before the lognormal model is used. The lognormal distribution has pdf:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\mu}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) \qquad \text{for } x > 0 \tag{2.5}$$

Some authors (Evans et al., 2012; Radicchi et al., 2008) have used the lognormal model to fit citation count data to cited articles only but others (Thelwall, 2016c) included uncited articles by adding one to all citation counts.

Nonetheless, since the lognormal is a continuous distribution, it is necessary to discretise the pdf before fitting it to discrete data, as the probability of obtaining any integer value in a continuous distribution is zero. Thelwall and Wilson

(2014b) simulated discretised lognormal data by rounding continuous real numbers to the nearest integer. Thelwall and Wilson (2014a); Thelwall (2016c) used the following discretisation method:

$$g(x;\mu,\sigma) = \frac{f(x;\mu,\sigma)}{\sum_{i=1}^{i=\infty} f(i;\mu,\sigma)} \qquad \text{for } x = 1, 2, \ldots \tag{2.6}$$

where $f(x;\mu,\sigma)$ is as at Equation 2.5.

2.2.6 Poisson models

The Poisson distribution is the most commonly used count distribution. It has only one parameter and it assumes an equal mean-variance relationship. The Poisson pmf is:

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{for } x = 0, 1, 2, \ldots \tag{2.7}$$

Vieira and Gomes (2010) used the Poisson, exponential-Poisson mixture ($\lambda = \frac{1}{E}\exp(-\lambda/E)$), and double exponential-Poisson models (where λ is obtained by the average of two exponentials) to model citation counts, and found that the double exponential-Poisson fitted citation distributions examined well.

2.2.7 Negative binomial models

The negative binomial distribution is a two parameter distribution. The first parameter is the mean parameter and the second parameter is the size or dispersion parameter. There are various parameterisations used for the negative binomial model, including NB1, NB2, NB-H and NB-P, where each accounts for a different variance function (Hilbe, 2011, p.285). The two most commonly used are NB1 and NB2.

The NB1 model has pmf:

$$f(y;\mu,\alpha) = \frac{\Gamma(y + (\mu/\alpha))\alpha^y}{\Gamma(\mu/\alpha)\Gamma(y+1)(1+\alpha)^{y+(\mu/\alpha)}} \qquad \text{for } y = 0, 1, 2, \ldots \tag{2.8}$$

and mean/variance relationship $\sigma^2 = \mu + \mu\alpha$.

On the other hand, the NB2 model has pmf:

$$f(y;\mu,\alpha) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)} \cdot \frac{(\mu\alpha)^y}{(\mu\alpha+1)^{y+(1/\alpha)}} \qquad \text{for } y = 0, 1, 2, \ldots \tag{2.9}$$

and mean/variance relationship $\sigma^2 = \mu + \mu^2\alpha$. Note that this NB2 pmf may also be written as:

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \cdot \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \cdot \left(1-\frac{\theta}{\mu+\theta}\right)^{y} \qquad \text{for } y = 0,1,2,\ldots \tag{2.10}$$

with mean/variance relationship $\sigma^2 = \mu + \mu^2/\theta$. Here θ is a reciprocal of α in (2.9).

The negative binomial model is a Poisson-gamma mixture, that is, a Poisson distribution with parameter λ, where λ is itself a random variable, following a Gamma distribution with scale parameter α and shape parameter β:

$$x \sim \mathit{Poisson}(\lambda) \;\Rightarrow f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2,\ldots \tag{2.11}$$

$$\lambda \sim \mathit{Gamma}(\alpha,\beta) \;\Rightarrow g(\lambda;\alpha,\beta) = \frac{\beta^{\alpha}\lambda^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda\beta} \qquad \lambda > 0 \tag{2.12}$$

The unconditional distribution of y is obtained using the joint density of x and λ, that is:

$$f(x;\lambda,\alpha,\beta) = \int_0^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!}\frac{\beta^{\alpha}\lambda^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda\beta}\, d\lambda \tag{2.13}$$

Solving this will result in Equation 2.9 (Hilbe, 2011, p.188-189).

2.2.8 Poisson inverse Gaussian models

The Poisson inverse Gaussian (PIG) models can be used as an alternative to the negative binomial model in the presence of over-dispersed data (Willmot, 1987). The PIG model has pmf:

$$f(y|\mu,\sigma) = \sqrt{\frac{2\alpha}{\pi}}\mu^{y}e^{\frac{1}{\sigma}}\frac{K_{y-\frac{1}{2}}(\alpha)}{(\alpha\sigma)^{y}y!} \qquad y = 0,1,2,\ldots \tag{2.14}$$

where $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$; $\mu > 0$; $\sigma > 0$; $K_{\lambda}(t) = \frac{1}{2}\int_0^{\infty} x^{\lambda-1}e^{\left(-\frac{1}{2}t(x+x^{-1})\right)}dx$, and K_{λ} is also known as the modified Bessel function of the third kind (Dean et al., 1989; Rigby et al., 2005).

This is also known as the inverse Gaussian-Poisson distribution, which was introduced by Sichel (1974) to model sentence length.

2.2.9 Hurdle models

In the case of excess zeros in the model, that is, having a greater number of zeros than expected from a given model, a hurdle or zero-inflated model is often used (Mullahy, 1997). A hurdle model (also known as a two-part model (Heilbron,

1994)) consists of two components. The first models the zero state, while the second component models the non-zero state, which are usually positive counts (Mullahy, 1986). The second component is often known as zero-truncated. The idea of a hurdle comes from the need for the observed value to cross the zero hurdle, before being modelled (Zuur et al., 2013). Cameron and Trivedi (1999) used this model to describe the number of visits to a doctor, where whether a person visits a doctor for the first time is based on his/her choice, and this is classed as a 'hurdle' or the zero-state. After the first visit, the person goes to the non-zero state, where he or she may need to revisit the doctor again depending on whether their health condition improves.

If the probability of observing a zero is π, and $g(x)$ is the distribution function of the non-zero state, then the pmf of a hurdle model is:

$$f(x;\pi,\Theta) = \begin{cases} \pi & \text{if } x = 0 \\ \frac{1-\pi}{1-g(0;\Theta)} \cdot g(x;\Theta|x>0) & \text{if } x > 0 \end{cases} \tag{2.15}$$

(Cameron and Trivedi, 1999). The Poisson and negative binomial distributions are frequently used for the non-zero state. In general, Ridout et al. (1998) noted that it is possible to use any distribution that will suitably model the non-zero state, such as the logarithmic distribution.

Didegah and Thelwall (2013b) used the negative binomial hurdle models in citation count analysis and found that they fitted better than the standard negative binomial and zero-inflated negative binomial models in terms of AIC (see Section 2.4.2 for details about AIC).

2.2.10 Zero-inflated models

Like the hurdle models, zero-inflated models are commonly used in practice when the data contain a large number of zeros. However, this is applicable even if few zeros are present. This is because it depends on the number of zeros expected by a fitted model. Hence, the presence of one zero where an occurrence is not expected will also lead to zero-inflation.

Zero-inflated models classify the zeros into two categories, the 'perfect zeros' (also known as structural zeros) and the zeros observed as a result of the distribution of interest (also known as count zeros) (Lambert, 1992). A 'perfect zero' sometimes refers to an event that is impossible. An example is when modelling the number of pregnant people in a room, men will form the 'perfect zeros' as it is naturally impossible to have a pregnant man.

In the example used by Zuur et al. (2009) when observing the number of hippopotami, a 'perfect zero' could be due to observer error, whilst a count zero

is obtained simply because no hippopotamus is present in that habitat. Zuur et al. (2009) used the terms 'true' and 'false' zeros to represent count and perfect zeros respectively but this is context specific, as in this case it is based on observations made by scientists in a habitat.

The pmf of a zero-inflated model is:

$$f(x;\pi,\Theta) = \begin{cases} \pi + (1-\pi)f(0;\Theta) & x = 0 \\ (1-\pi)f(x;\Theta) & x > 0 \end{cases} \tag{2.16}$$

where $f(x)$ is the probability function of the distribution of interest and π is the probability of obtaining a perfect zero.

Didegah and Thelwall (2013a) investigated the suitability of zero-inflated models in citation analysis but no interpretation of perfect zero was given. Thelwall (2016a) investigated the use of zero inflated discretised lognormal and zero inflated hooked power law in modelling citation counts by interpreting perfect zeros in this context as uncitable articles, such as those in academic related magazines available in the Scopus database. Some examples given include news related industry-focused magazines in pharmaceutical science such as 'Pharmaceutisch Weekblad' and intellectual magazines in cultural studies such as 'North American Review'.

2.3 Standard compound distributions

Compound distributions (also known as generalised and stopped sums distributions (Johnson et al., 2005)) are discrete distributions that are used in a wide range of applications, such as branching processes in ecology (Foster and Bravington, 2013), damage processes (O'Keeffe et al., 2013), and risk assessments (Zhang et al., 2014).

Compound Poisson models are commonly used to address over-dispersion in data (Cox, 1983). Over-dispersion occurs when the variance is greater than the mean relative to a distribution. In this thesis, the term over-dispersion is relative to the Poisson distribution, that is, when the Poisson variance is greater than the relative Poisson mean. A compound Poisson model may be interpreted as the random sum of independent random Poisson variables:

$$S_N = X_1 + X_2 + X_3 + \cdots + X_N \tag{2.17}$$

where the number of terms N in the sum is Poisson distributed (Daley and Vere-Jones, 2003; Zhang et al., 2014). It is clear that for $N = 0$, $S_N = 0$.

Previous research has used epidemiological scenarios to mimic the spread of ideas in academia (Goffman and Newill, 1964). Using a similar approach, in this thesis, we investigate several compound distributions which are commonly used in branching processes.

2.3.1 Neyman type A

The Neyman type A distribution is a compound Poisson-Poisson distribution, which was first introduced by Neyman (1939) to model the number of larvae in a field. Neyman first assumed that the number of masses of eggs laid by moths, indexed by i, follows a Poisson distribution, and that the number of egg masses equals N. Each individual egg mass i (first generation) then independently hatches and gives rise to X_i larvae (second generation), which also follows a Poisson distribution.

Suppose that the first generation has Poisson parameter λ and the second generation has Poisson parameter ϕ, then the Neyman type A distribution has pmf:

$$f(x;\lambda,\phi) = \frac{\mathrm{e}^{-\lambda}\phi^x}{x!}\sum_{j=0}^{\infty}\frac{\left(\lambda \mathrm{e}^{-\phi}\right)^j j^x}{j!} \tag{2.18}$$

(Johnson et al., 2005). The pmf of the Neyman type A may also be written as a recurrence relation, using its mean μ and dispersion index δ as parameters (Johnson et al., 2005; Oliveira et al., 2016):

$$f(x;\mu,\delta) = \exp(-\lambda + \exp(-\phi)\lambda)\frac{\phi^x \mu_x(\exp(-\phi)\lambda)}{x!} \qquad x > 0 \tag{2.19}$$

where

$$\mu_x = \sum_{k=0}^{x} R(x,k)\theta^k \qquad ; \qquad R(x,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^x \qquad ;$$

$$\phi = \delta - 1 \qquad \text{and} \qquad \lambda = \frac{\mu}{\delta - 1}$$

Dobbie and Welsh (2001) incorporated covariates into the Neyman type A model and suggest the appropriateness of using the Neyman type A model in contagious or clustered data.

2.3.2 Polya-Aeppli

The Polya-Aeppli distribution is a compound Poisson-shifted geometric distribution, with pmf:

$$f(x;\theta,p) = \begin{cases} e^{-\theta} \\ e^{-\theta}p^x \sum_{j=1}^{x} \binom{x-1}{j-1} \frac{(\theta q/p)^j}{j!} \end{cases} \tag{2.20}$$

This was first used for clusters of objects, where the number of clusters are assumed to follow a Poisson distribution, and the number of objects per cluster follows a shifted geometric distribution (Johnson et al., 2005, p.410-411).

2.3.3 Delaporte

The Delaporte distribution is a special case of the Poisson shifted generalised inverse Gaussian distribution (Rigby et al., 2008), also known as a convolution of Poisson and negative binomial distribution in the actuary field or Lüders Formel II distribution (Lüders, 1934). The Delaporte distribution has three parameters, λ (the Poisson parameter), α and β (shape and scale parameter), and pmf:

$$f(x;\lambda,\alpha,\beta) = \sum_{i=0}^{x} \frac{\Gamma(\alpha+i)}{\Gamma(\alpha)i!(n-i)!} \cdot \frac{\beta^i \lambda^{x-i} \mathrm{e}^{-\lambda}}{(1+\beta)^{\alpha+i}} \tag{2.21}$$

where $x = 0, 1, 2, \cdots$ and $\lambda, \alpha, \beta > 0$ (Adler, 2014).

2.3.4 Other compound models considered

In this thesis, various standard compound models have been considered to allow comparisons with the variant models developed. The models considered are:

(i) Compound Poisson-NB

(ii) Compound NB-Poisson

(iii) Compound NB-NB

The compound Poisson-NB is also known as the Poisson-Pascal or generalised Polya-Aeppli distribution and was first introduced by Skellam (1952).

Some properties of these models, such as their probability generating functions (pgf) are explored. For any compound A-B distribution, which can be viewed as a random sum, its pgf $G(z)$ can be obtained using $G(z) = G_1(G_2(z))$, where G_1 is the pgf of distribution A, and G_2 is the pgf of distribution B (Johnson et al.,

2005, p.361). Hence, the pgf of the compound Poisson-NB distribution is:

$$\begin{aligned} G_1(z) &= \exp(\lambda(z-1)) && (2.22)\\ G_2(z) &= \left(\frac{p}{1-z+zp}\right)^{\alpha} && (2.23)\\ \therefore G(z) &= \exp\left(\lambda\left(\left(\frac{p}{1-z+zp}\right)^{\alpha}-1\right)\right) && (2.24)\\ &= \exp(-\lambda)\cdot\exp\left(\lambda\left(\frac{p}{1-z+zp}\right)^{\alpha}\right) && (2.25)\\ &= \exp(-\lambda)\cdot\sum_{j=0}^{\infty}\frac{1}{j!}\left(\lambda\left(\frac{p}{1-z+zp}\right)^{\alpha}\right)^{j} && (2.26) \end{aligned}$$

Using similar method, the pgf of the other compound models considered are:

(i) Compound NB-Poisson distribution(μ, α, λ)
Let $p = \frac{\alpha}{\mu+\alpha}$,

$$G(z) = \left(\frac{p}{1-\exp(\lambda(z-1))+p\exp(\lambda(z-1))}\right)^{\alpha} \quad (2.27)$$

(ii) Compound NB-NB distribution$(\mu_1, \alpha_1, \mu_2, \alpha_2)$
Let $Z = \left(\frac{q}{qt-t+1}\right)^{\theta}$, $p = \frac{\alpha_1}{\mu_1+\alpha_1}$ and $q = \frac{\alpha_2}{\mu_2+\alpha_2}$,

$$G(z) = \left(\frac{p}{pZ-Z+1}\right)^{\alpha} \quad (2.28)$$

It is possible to obtain the probability mass function (pmf) of these models from the pgf by:

$$f(x;\Theta) = \frac{G^n(0;\Theta)}{n!} \quad (2.29)$$

However, for these models, the pmf obtained by this procedure are extremely long and are not required. Thus, they are not reported here. The expectations and variances of these models are given in Appendix B.

2.4 Model selection and validation techniques

2.4.1 Maximum likelihood estimation

Let $Y_1, Y_2, \cdots, Y_n$ be n independent random variables with pdf/pmf $f(y_i;\Theta)$. Given a vector of parameters Θ, the likelihood function of the data, $\{y_i\}$, is:

$$L(\Theta; y) = \Pi_{i=1}^{n} f(y_i;\Theta) \quad (2.30)$$

This is the probability that the data, $\{y_i\}$, is from the distribution $f(\Theta)$, hence we aim to maximise this probability, which is also the likelihood function, $L(\Theta; u)$. The value of the vector of parameters, Θ_*, which maximises Equation 2.30 is the maximum likelihood estimator and this process is known as maximum likelihood estimation. Given that the numerical value of $L(\Theta_*; y)$ is usually very close to zero, it is more convenient to use its natural log. This is commonly known as the log-likelihood function:

$$\log L(\Theta; y) = \sum_{i=1}^{n} \log \left(f(y_i; \Theta) \right) \tag{2.31}$$

2.4.2 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) was proposed by Akaike (1973) and is a commonly used model selection criterion. AIC is derived from Kullback-Leibler (K-L) distance (Kullback and Leibler, 1951), which measures information lost when the true model is replaced by an approximation (Burnham and Anderson, 2004). If the true model is f and we use model g to approximate f, then for the continuous case, the K-L distance, $I(f, g)$ is:

$$\begin{aligned} I(f,g) &= \int f(x) \log \left(\frac{f(x)}{g(x;\theta)} \right) dx & (2.32) \\ &= \int f(x) \log(f(x)) dx - \int f(x) \log(g(x;\theta)) dx & (2.33) \\ &= E_f \left(\log(f(x)) \right) - E_f \left(\log(g(x;\theta)) \right) & (2.34) \end{aligned}$$

Since $I(f, g)$ measures information lost, a smaller $I(f, g)$ indicates a better model. Although f is not known, in Equation 2.34, the term $E_f \left(\log(f(x)) \right)$ is a constant, which does not depend on sample size, n or parameters, θ. As the term suggests, Equation 2.34 shows that the K-L distance can be interpreted as the 'distance' between models f and g. Akaike (1973) found that maximising the log-likelihood is equivalent to a biased estimator, which is approximately equal to the number of parameters used in the chosen model, k. Thus, the relative K-L distance is:

$$\log L(\hat{\theta}) - k \tag{2.35}$$

Here, k is the asymptotic bias correction term (Burnham and Anderson, 2004). Akaike (1973) multiplies Equation 2.35 by -2 to give the Akaike Information Criterion (AIC):

$$AIC = -2 \log L(\hat{\theta}) + 2k \tag{2.36}$$

Although the choice of -2 is not explicitly justified by Akaike (1973), DeLeeuw (1992) stated that this is somewhat mysterious, while others (Burnham and Anderson, 2002, p.64) concluded that this could be due to historical reasons as doing so will give an asymptotically chi-squared value under certain conditions. Vrieze (2012) stated that AIC may not be consistent in selecting the true model and AIC may be less efficient if the true model is not within the selection. Hurvich and Tsai (1989, 1995) proposed a bias correction to the AIC for small samples in normal regression and autoregressive models. They denote this AIC_c, which includes an extra sample size penalty:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \tag{2.37}$$

2.4.3 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC), also known as the Schwarz criterion (SIC or SBC), was derived by Schwarz (1978) using a Bayesian procedure and gives a rough approximation to the logarithm of the Bayes factor (Kass and Raftery, 1995), where a larger posterior probability of a candidate model indicates a better model.

Let $Y = Y_1, Y_2, \cdots, Y_n$ be n independent random variables and suppose we wish to test two hypothesis H_1 and H_2, then by Bayes theorem,

$$\frac{P(H_1|Y)}{P(H_2|Y)} = \frac{P(Y|H_1)}{P(Y|H_2)} \times \frac{P(H_1)}{P(H_2)} \tag{2.38}$$

which is equivalent to:

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds} \tag{2.39}$$

In the simplest case, if H_1 and H_2 are single distributions with no free parameters, then the Bayes factor is equivalent to the likelihood ratio, where for each hypothesis, k:

$$P(Y|H_k) = \int P(Y|\theta_k, H_k) \cdot \pi(\theta_k|H_k)\, d\theta_k \tag{2.40}$$

where θ_k is the parameter under H_k, $\pi(\theta_k|H_k)$ is the prior density and $P(Y|\theta_k, H_k)$ is the probability density of Y given the value of θ_k (Kass and Raftery, 1995). The advantage of the Schwarz criterion is that it does not require the prior distributions, $\pi(\theta_k|H_k)$. If θ_* is the maximum likelihood estimator H, d is the dimension of θ and n is the sample size, then the Schwarz criterion uses:

$$S = \log(P(Y|\theta_{1*}, H_1)) - \log(P(Y|\theta_{2*}, H_2)) - \frac{1}{2}(d_1 - d_2)\log(n) \tag{2.41}$$

If we let the Bayes factor of H_1 against H_2 be B_{12}, Kass and Raftery (1995) also stated that:

$$\text{As } n \to \infty, \qquad \frac{S - \log B_{12}}{\log B_{12}} \to 0 \tag{2.42}$$

and multiplying the Schwarz criterion by -2 gives the BIC:

$$BIC = -2 \log L(\hat{\theta}) + k \log(n) \tag{2.43}$$

such that k is the number of parameters and n is the number of observations.

AIC versus BIC

Comparing Equations 2.36 and 2.43, BIC and AIC will only differ greatly when there is a large number of observations. More specifically, BIC has a stronger penalty on the number of parameters for $n \geq 8$ (as $\log 8 \approx 2$) (Claeskens and Hjort, 2008, p.70). Criteria can be classified as *consistent* or *efficient* as follows. Assuming that a true model exists in the set of models used, a consistent criterion is one which will select the true model with probability close to one as n increases, but an efficient criterion will select the model which minimises the mean squared error of prediction. It is known that AIC is asymptotically efficient but not consistent, whereas BIC is not efficient but consistent (Claeskens and Hjort, 2008).

Kass and Raftery (1995) concluded that AIC tends to select models with a large number of parameters when n is large (known as overfitting) but BIC favours simpler models. However, BIC has the tendency to underfit when n is small (Dziak et al., 2012). When comparing models with similar AIC/BIC, the simpler model (also known as the parsimonious model), that is one with fewer number of parameters, is often chosen (Claeskens and Hjort, 2008). For cases when they disagree, that is when a different model is preferred by each criterion, judgement should be made based on theory. Dziak et al. (2012) suggest to use the BIC favoured model as a minimum (simplest model) and the AIC favoured model as a maximum (model with most number of parameters). Kadane and Lazar (2004) suggest using these criteria to eliminate clearly inappropriate models instead of selecting the 'best' model.

In citation analysis, the number of observations is often very large. For example, if $n = 5000$, an increase in k will lead to an increase of 8.52 (as $\log(5000) = 8.52$) in the BIC, but only an increase of 2 in AIC. Hence, perhaps the BIC should be used in citation analysis to avoid over fitting. Nonetheless, it is difficult to pinpoint which criteria to use (Dziak et al., 2012) and conclusions should not be made solely based on a single criterion. Einbeck and Wilson (2016) showed that a model favoured by AIC/BIC may still be too flexible, resulting in

fits that are too good to be true. In addition, given that there are many factors that affect citations, such as journal impact factor, international collaboration and abstract length (Didegah and Thelwall, 2013b), obtaining a 'true' model across all disciplines in citation analysis is not a realistic goal as citation patterns are known to vary across disciplines.

2.4.4 Standard errors of parameter estimates

The standard errors of the parameter estimates can be calculated using the Hessian matrix, which is a square matrix of the second order partial derivatives of the log-likelihood function. Suppose l represents the log-likelihood function of a distribution with two parameters, λ_1 and λ_2, then the Hessian matrix is:

$$\begin{bmatrix} \frac{\partial^2 l}{\partial \lambda_1^2} & \frac{\partial^2 l}{\partial \lambda_1 \partial \lambda_2} \\ \frac{\partial^2 l}{\partial \lambda_1 \partial \lambda_2} & \frac{\partial^2 l}{\partial \lambda_2^2} \end{bmatrix} \tag{2.44}$$

The above 2×2 matrix can be extended based on the number of parameters. For example, the Hessian matrix of a function with 5 parameters will be a 5×5 matrix. If the estimated parameters, $\hat{\lambda_i}$, where $i = 1, 2, \cdots, n$, in (2.44) are the maximum likelihood estimators, then the negative of this is equivalent to the observed information matrix. Since the inverse of the observed information matrix is the asymptotic covariance matrix of $\hat{\lambda_i}$, the i^{th} diagonal element will be the variance of $\hat{\lambda_i}$. Hence, the standard errors of parameters $\lambda_1, \lambda_2, \cdots \lambda_n$ are the square root of the main diagonal of the inverse of the negative Hessian matrix (Hogg and Craig, 1995, p.384-385).

2.4.5 Randomised quantile residuals

Dunn and Smyth (1996) introduced randomised quantile residuals for model checking, where the fitted distribution function is inverted at each response value to find its equivalent standard normal quantile.

For the continuous cumulative distribution function $F(y_i; \mu, \phi)$, the randomised quantile residual for y_i is defined by:

$$r_{q,i} = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})) \tag{2.45}$$

where $\Phi()$ is the cumulative distribution of the standard normal distribution and $F(y_i; \hat{\mu}_i, \hat{\phi})$ is uniformly distributed on the unit interval (Dunn and Smyth, 1996).

This method introduced by Dunn and Smyth (1996) also allows the computation of continuous residuals from discrete responses. For the discrete cumulative distribution function $F(y_i; \mu, \phi)$, the randomised quantile residual for y_i is defined

by:

$$r_{q,i} = \Phi^{-1}(u_i) \tag{2.46}$$

where $\Phi()$ is the cumulative distribution of the standard normal distribution, and u_i is a uniform random variable on the interval $(a_i, b_i]$ (Dunn and Smyth, 1996). Although a_i is defined as $\lim_{y \uparrow y_i} F(y_i; \hat{\mu}_i, \hat{\phi})$, for simplicity and following Smyth et al. (2015), we obtain the intervals a_i and b_i using $F(y_i; \mu, \phi)$, such that $a_i = F(y_i - 1; \mu, \phi)$ and $b_i = F(y_i; \mu, \phi)$.

Chapter 3

Variants of compound models

3.1 Introduction

Recall that the standard compound model can be viewed as two generations, where the second is a consequence of the first. For example using the analogy of aphids, the two generations could be viewed as winged parent aphids (first generation), which fly in to a plot and give birth to their wingless offspring (second generation), which will remain in situ. Here, the numbers of both parent and offspring aphids follow a statistical distribution individually, and the compound model will only model the total number of second generation offspring.

In this chapter, two variant models are introduced, in Sections 3.2 and 3.3. Both models may be viewed as a sum of two generations and cases where both generations are negative binomial distributed, or one is Poisson and the other is negative binomial are considered. In Section 3.2, the first variant, SVB, which is also equivalent to a convolution model is discussed. In Section 3.3, a second variant, denoted SVA, is introduced. The notations SVA and SVB are used to represent "stopped sum variant" A and B. This is because it is possible to regard the SVA and SVB models as variants of compound (also known as stopped sum) models (see Section 3.3.1).

We describe in detail the properties of these distributions in Section 3.4. We show how the standard compound models and their variants are modelled in the R software system in Section 3.5. In Section 3.6, we explain the methods used to estimate parameters in the proposed models.

3.2 SVB distributions

We first investigate some convolutions of Poisson and negative binomial distributions. In this thesis, we denote the convolution models as SVB models.

For example, if X and Y are two independent count random variables where $X \sim Poisson(\lambda)$ and $Y \sim NB(\mu, \alpha)$, then we denote their convolution as a SVB Poisson-NB distribution. Note that the SVB Poisson-NB is also known as the Lüders Formel II distribution (Lüders, 1934), or in actuarial science the Delaporte distribution (Johnson et al., 2005, p.242). The pmf of the SVB Poisson-NB distribution may be written as:

$$f(x; \lambda, \mu, \alpha) = \sum_{j=0}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!} \binom{x-j+\alpha-1}{\alpha-1} p^\alpha q^{x-j} \qquad x = 0, 1, 2, \ldots \quad (3.1)$$

where $p = \alpha/(\mu + \alpha)$ and $q = 1 - p$.

The other SVB distribution investigated is the SVB NB-NB distribution, with pmf:

$$f(x; \mu_1, \alpha, \mu_2, \theta) = \sum_{j=0}^{x} \binom{j+\alpha-1}{\alpha-1} p^\alpha q^j \binom{x-j+\theta-1}{\theta-1} r^\theta s^{x-j} \qquad x = 0, 1, 2, \ldots \quad (3.2)$$

where $p = \alpha/(\mu_1 + \alpha)$, $q = 1 - p$, $r = \theta/(\mu_2 + \theta)$ and $s = 1 - r$.

Since the convolution of two Poisson models is equivalent to a Poisson model, it is not considered. The convolution of two negative binomial distributions is a negative binomial distribution only if both NB generations have equal values of the size parameter, that is:

$$X * Y \sim NB(\mu_1 + \mu_2, \alpha)$$
$$\text{if and only if } X \sim NB(\mu_1, \alpha) \text{ and } Y \sim NB(\mu_2, \alpha) \quad (3.3)$$

where the convolution of distributions X and Y may be denoted as $X * Y$. However, this is not necessarily the case with the SVB models proposed.

Here, the order of convolution of two generations is unimportant thus the moment generating function of the convolution is the product of the two generating functions. For example, in general SVB D1-D2 and SVB D2-D1 are the same, where D1 and D2 are any two distributions (Baglivo, 2005). As SVB Poisson-NB and SVB NB-Poisson are equivalent (see proof in (3.17) in Section 3.4.1), only SVB Poisson-NB is used in this thesis.

3.3 SVA distributions

We present a second variant, denoted SVA, which follow a zero restriction, that is, a zero in the first generation will automatically be followed by a zero in the second generation. In general, if D_1 and D_2 are discrete probability distributions

with pmfs f_{D1} and f_{D2} respectively, then a SVA D1-D2 distribution will have pmf:

$$f(x) = \begin{cases} f_{D1}(0) & x = 0 \\ \sum_{j=1}^{x} \left(f_{D1}(j) \times f_{D2}(x-j) \right) & x = 1, 2, 3, \cdots \end{cases} \tag{3.4}$$

If the first generation follows a Poisson distribution with parameter λ and the second generation follows a negative binomial distribution with parameters μ and α, then this SVA Poisson-NB distribution has pmf:

$$f(x; \lambda, \mu, \alpha) = \begin{cases} \mathrm{e}^{-\lambda} & x = 0 \\ \sum_{j=1}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^{j}}{j!} \binom{x-j+\alpha-1}{\alpha-1} p^{\alpha} q^{x-j} & x = 1, 2, 3, \cdots \end{cases} \tag{3.5}$$

where $p = \alpha/(\mu + \alpha)$ and $q = 1 - p$.

Using similar principles, if the first generation follows a negative binomial distribution with parameters μ and α, where $p = \alpha/(\mu + \alpha)$ and $q = 1 - p$, and the second generation follows a Poisson distribution with parameter λ, then this SVA NB-Poisson distribution has pmf:

$$f(x; \mu, \alpha, \lambda) = \begin{cases} p^{\alpha} & x = 0 \\ \sum_{j=1}^{x} \binom{j+\alpha-1}{\alpha-1} p^{\alpha} q^{j} \frac{\mathrm{e}^{-\lambda}\lambda^{x-j}}{(x-j)!} & x = 1, 2, 3, \cdots \end{cases} \tag{3.6}$$

The other case considered is SVA NB-NB distribution, where the first generation follows a negative binomial distribution with parameters μ_1 and α, while the second generation follows a negative binomial distribution with parameters μ_2 and θ, then this SVA NB-NB distribution has pmf:

$$f(x; \mu_1, \alpha, \mu_2, \theta) = \begin{cases} p^{\alpha} & x = 0 \\ \sum_{j=1}^{x} \binom{j+\alpha-1}{\alpha-1} p^{\alpha} q^{j} \binom{x-j+\theta-1}{\theta-1} r^{\theta} s^{x-j} & x = 1, 2, 3, \cdots \end{cases} \tag{3.7}$$

where $p = \alpha/(\mu_1 + \alpha)$, $q = 1 - p$, $r = \theta/(\mu_2 + \theta)$ and $s = 1 - r$.

Here, unlike SVB, in general the order of the distributions within the SVA distributions does matter.

3.3.1 Alternative interpretation of SVA and SVB distributions

The notations SVA and SVB were initially used to represent "stopped sum variant A" and "stopped sum variant B". This is because, alternatively, they may be viewed as variants of compound (also known as stopped sum) distributions. Given a compound model of the form

$$S_N = X_1 + X_2 + \cdots + X_N \tag{3.8}$$

where N represents the first generation and S_N represent the second generation, then the SVA distribution is the sum of the two generations, $S_N + N$. It is clear here that if $N = 0$, $S_N = 0$. Consequently, the SVB distribution is a variant that does not possess this zero restriction, that is, $S_N \geq 0$ when $N = 0$.

3.4 Properties of SVA and SVB distributions

3.4.1 Moment generating functions of SVA and SVB distributions

The moment generating function (mgf) of a random variable X is:

$$M[X;t] = E[\mathrm{e}^{tX}] \tag{3.9}$$

$$= \sum_x \mathrm{e}^{tx} \cdot P(X = x) \qquad \text{if } X \text{ is discrete; or} \tag{3.10}$$

$$= \int_{-\infty}^{\infty} \mathrm{e}^{tx} \cdot f(x)dx \qquad \text{if } X \text{ is continuous} \tag{3.11}$$

(Kinney, 1997, p.267). Note that not all distributions have mgfs, for example the Cauchy distribution. This is because the Cauchy distribution has infinite moments and thus the expected value does not exist (Grimmett and Welsh, 1986, p.112).

Since the SVB distributions are convolutions, we use the convolution theorem to obtain their moment generating functions. The convolution theorem states that, if X and Y are independent random variables, such that $Z = X + Y$, where

$$P(Z = z) = \sum_{k=-\infty}^{\infty} P(X = k)\ P(Y = z - k) \tag{3.12}$$

(Baglivo, 2005, p.64), then Z has mgf:

$$M(Z,t) = M(X,t)\ M(Y,t) \tag{3.13}$$

Moment generating functions of SVB distributions

Since our SVB distributions are convolutions of Poisson and negative binomial distributions, we first need to obtain the mgf of the Poisson and negative binomial distributions. In general, the Poisson distribution with parameter λ has mgf:

$$\begin{aligned} M_{Pois}(t) =& \sum_{j=0}^{\infty} \mathrm{e}^{tj}\ P(X=j) \\ =& \sum_{j=0}^{\infty} \mathrm{e}^{tj}\ \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!} \\ =& \mathrm{e}^{-\lambda} \sum_{j=0}^{\infty} \frac{(\mathrm{e}^t\lambda)^j}{j!} \\ =& \ \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda\mathrm{e}^t} \\ =& \ \mathrm{e}^{\lambda(\mathrm{e}^t-1)} \end{aligned} \tag{3.14}$$

and the negative binomial (NB2) distribution with parameters μ and α, where

$p = \alpha/(\mu+\alpha)$ and $\sigma^2 = \mu + \mu^2/\alpha$ has mgf:

$$\begin{aligned}
M_{NB}(t) =& \sum_{x=0}^{\infty} \mathrm{e}^{tx} \binom{x+\alpha-1}{\alpha-1} p^{\alpha}(1-p)^{x} \\
=& p^{\alpha} \sum_{x=0}^{\infty} \binom{x+\alpha-1}{\alpha-1} \left(\mathrm{e}^{t}(1-p)\right)^{x} \\
=& p^{\alpha} \sum_{x=0}^{\infty} \binom{x+\alpha-1}{x} \left(\mathrm{e}^{t}(1-p)\right)^{x} \\
=& p^{\alpha} \sum_{x=0}^{\infty} (-1)^{x} \binom{-\alpha}{x} \left(\mathrm{e}^{t}(1-p)\right)^{x} \\
=& p^{\alpha} \sum_{x=0}^{\infty} \binom{-\alpha}{x} \left(-\mathrm{e}^{t}(1-p)\right)^{x} \\
=& p^{\alpha} \left(1-(1-p)\mathrm{e}^{t}\right)^{-\alpha} \\
=& \left(\frac{p}{1-(1-p)\mathrm{e}^{t}}\right)^{\alpha}
\end{aligned} \tag{3.15}$$

Consequently, the mgf of the SVB distributions computed using (3.13) are:

(i) SVB Poisson-NB

$$M(X,t) = \mathrm{e}^{\lambda(\mathrm{e}^{t}-1)} \left(\frac{p}{1-(1-p)\mathrm{e}^{t}}\right)^{\alpha} \qquad \text{where } p = \frac{\alpha}{\mu+\alpha} \tag{3.16}$$

(ii) SVB NB-Poisson

$$M(X,t) = \mathrm{e}^{\lambda(\mathrm{e}^{t}-1)} \left(\frac{p}{1-(1-p)\mathrm{e}^{t}}\right)^{\alpha} \qquad \text{where } p = \frac{\alpha}{\mu+\alpha} \tag{3.17}$$

Note that this is the same as the mgf of SVB Poisson-NB in (3.16).

(iii) SVB NB-NB

$$M(X,t) = \left(\frac{p}{1-(1-p)\mathrm{e}^{t}}\right)^{\alpha} \left(\frac{r}{1-(1-r)\mathrm{e}^{t}}\right)^{\theta} \tag{3.18}$$

Moment generating functions of SVA distributions

The mgfs of SVA distributions can be computed indirectly, using those of the SVB distributions. Let X and Y be random variables where $X \sim SVA$ and $Y \sim SVB$, then we can write the SVA distribution as:

$$f_{SVA}(n) = \begin{cases} a_0 & n = 0 \\ \sum_{i=1}^{n} a_i b_{n-i} = a_1 b_{n-1} + a_2 b_{n-2} + ... + a_n b_0 & n \neq 0 \end{cases} \tag{3.19}$$

where a and b represent the counts from the first and second generations with distributions A and B respectively. In general, the SVB distribution is:

$$\begin{aligned} f_{SVB}(n) =& \sum_{i=0}^{n} a_i b_{n-i} \\ =& a_0 b_n + a_1 b_{n-1} + ... + a_n b_0 \end{aligned} \tag{3.20}$$

So for $n \neq 0$, $f_{SVA}(n) = f_{SVB}(n) - a_0 b_n$

Therefore, the mgf of SVA distributions can be derived as follows:

$$\begin{aligned}
M(X,t) =& E(\mathrm{e}^{xt}) \\
=& \sum_{x=0}^{\infty} P(X=x)\mathrm{e}^{tx} \\
=& P(X=0) + \sum_{x=1}^{\infty} P(X=x)\mathrm{e}^{tx} \\
=& P(X=0) + \sum_{x=1}^{\infty} \Big(P(Y=x) - a_0 b_x\Big)\mathrm{e}^{tx} \\
=& P(X=0) + \sum_{x=1}^{\infty} P(Y=x)\mathrm{e}^{tx} - a_0 \sum_{x=1}^{\infty} b_x \mathrm{e}^{tx} \\
=& P(X=0) + \sum_{x=0}^{\infty} P(Y=x)\mathrm{e}^{tx} - P(Y=0) - a_0 \sum_{x=0}^{\infty} b_x \mathrm{e}^{tx} + a_0 b_0 \\
=& P(X=0) + E_Y(\mathrm{e}^{tx}) - a_0 E_B(\mathrm{e}^{tx}) \qquad \textit{since } P(Y=0) = a_0 b_0 \\
=& P(X=0) + M(Y,t) - a_0 M(B,t) \\
=& a_0 + M(Y,t) - a_0 M(B,t) \\
=& a_0(1 - M(B,t)) + M(Y,t)
\end{aligned} \tag{3.21}$$

where $M(B,t)$ is the moment generating function of the second generation distribution and $a_0 = f_{SVA}(0)$. Hence, the mgf of the SVA distributions considered are:

(i) SVA Poisson-NB

From (3.15), $M(B,t) = \left(\dfrac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha}$;

and from (3.16), $M(Y,t) = \mathrm{e}^{\lambda(\mathrm{e}^t-1)} \left(\dfrac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha}$

Thus, the mgf of SVA Poisson-NB is:

$$M(X,t) = \mathrm{e}^{-\lambda}\left(1 - \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha}\right) + \mathrm{e}^{\lambda(\mathrm{e}^t-1)} \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha} \tag{3.22}$$

(ii) SVA NB-Poisson

From (3.14), $M(B,t) = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}$;

and from (3.17), $M(Y,t) = \mathrm{e}^{\lambda(\mathrm{e}^t-1)} \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha}$

Thus, the mgf of SVA NB-Poisson is:

$$M(X,t) = p^{\alpha}\left(1-\mathrm{e}^{\lambda(\mathrm{e}^t-1)}\right) + \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha} \mathrm{e}^{\lambda(\mathrm{e}^t-1)} \tag{3.23}$$

(iii) SVA NB-NB

From (3.15), $M(B,t) = \left(\frac{r}{1-(1-r)\mathrm{e}^t}\right)^{\theta}$;

and from (3.18), $M(Y,t) = \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha} \left(\frac{r}{1-(1-r)\mathrm{e}^t}\right)^{\theta}$

Thus, the mgf of SVA NB-NB is:

$$M(X,t) = p^{\alpha}\left(1-\left(\frac{r}{1-(1-r)\mathrm{e}^t}\right)^{\theta}\right) + \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^{\alpha} \left(\frac{r}{1-(1-r)\mathrm{e}^t}\right)^{\theta} \tag{3.24}$$

3.4.2 Characteristic functions of SVA and SVB distributions

Suppose that the mgf exists for some discrete distribution, then we can obtain the characteristic function (cf) of the distribution by modifying its mgf, since the cf of a discrete distribution can be defined as:

$$\phi(t) = E[\mathrm{e}^{itX}] = \sum_{j=0}^{\infty} \mathrm{e}^{ijt} \cdot P(X=j) \tag{3.25}$$

(Johnson et al., 2005, p.57). Therefore, the characteristic functions, $\phi(t)$, of the SVA and SVB distributions are:

(i) SVA Poisson-NB

$$\phi(t) = \mathrm{e}^{-\lambda}\left(1 - \left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha}\right) + \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}\left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha} \tag{3.26}$$

(ii) SVA NB-Poisson

$$\phi(t) = p^{\alpha}\left(1 - \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}\right) + \left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha}\mathrm{e}^{\lambda(\mathrm{e}^{it}-1)} \tag{3.27}$$

(iii) SVA NB-NB

$$\phi(t) = p^{\alpha}\left(1 - \left(\frac{r}{1-(1-r)\mathrm{e}^{it}}\right)^{\theta}\right) + \left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha}\left(\frac{r}{1-(1-r)\mathrm{e}^{it}}\right)^{\theta} \tag{3.28}$$

(iv) SVB Poisson-NB

$$\phi(t) = \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}\left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha} \qquad \text{where } p = \frac{\alpha}{\mu+\alpha} \tag{3.29}$$

(v) SVB NB-NB

$$\phi(t) = \left(\frac{p}{1-(1-p)\mathrm{e}^{it}}\right)^{\alpha}\left(\frac{r}{1-(1-r)\mathrm{e}^{it}}\right)^{\theta} \tag{3.30}$$

3.4.3 Expectations and variances of SVA and SVB distributions

The expectation and variance of a distribution can be derived from their mgf:

$$M(t) = \sum_{x=0}^{\infty} e^{tx} P(X = x) \tag{3.31}$$

$$M'(t) = \sum_{x=0}^{\infty} x e^{tx} P(X = x) \tag{3.32}$$

$$\therefore M'(0) = \sum_{x=0}^{\infty} x P(X = x) \tag{3.33}$$

$$= E(X) \tag{3.34}$$

$$Var(X) = \sum_{x=0}^{\infty}(x-\mu)^2 P(X=x) \tag{3.35}$$

$$= \sum_{x=0}^{\infty}(x^2 - 2x\mu + \mu^2)P(X=x) \tag{3.36}$$

$$= \sum_{x=0}^{\infty} x^2 P(X=x) - 2\mu\sum_{x=0}^{\infty} xP(X=x) + \mu^2\sum_{x=0}^{\infty} P(X=x) \tag{3.37}$$

$$= \sum_{x=0}^{\infty} x^2 P(X=x) - 2\mu^2 + \mu^2 \tag{3.38}$$

$$= \sum_{x=0}^{\infty} x^2 P(X=x) - \mu^2 \tag{3.39}$$

$$= E(X^2) - (E(X))^2 \tag{3.40}$$

Since $$M''(t) = \sum_{x=0}^{\infty} x^2 e^{tx} P(X=x) \tag{3.41}$$

and $$M''(0) = \sum_{x=0}^{\infty} x^2 P(X=x) \tag{3.42}$$

$$\therefore Var(X) = M''(0) - (M'(0))^2 \tag{3.43}$$

The expectations and variances of the SVA and SVB distributions considered are given in Tables 3.1 and 3.2.

Table 3.1: *Expectations of the SVA and SVB distributions*

Distributions	*E(X)*
SVA Poisson-NB	$\lambda + \mu - \mathrm{e}^{-\lambda}\mu$
SVA NB-Poisson	$\lambda + \mu - \lambda\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}$
SVA NB-NB	$\mu_1 + \mu_2 - \mu_2\left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}$
SVB Poisson-NB	$\lambda + \mu$
SVB NB-NB	$\mu_1 + \mu_2$

Table 3.2: *Variances of the SVA and SVB models*

Distributions	*Var(X)*
SVA Poisson-NB	$E(X) + \frac{\mu^2}{\alpha} + \mathrm{e}^{-\lambda}\left(2\lambda\mu + \mu^2 - \mathrm{e}^{-\lambda}\mu^2 - \frac{\mu^2}{\alpha}\right)$
SVA NB-Poisson	$E(X) + \frac{\mu^2}{\alpha} + \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}\left[\lambda^2 + 2\lambda\mu - \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}\lambda^2\right]$
SVA NB-NB	$E(X) + \frac{\mu_1^2}{\alpha} + \frac{\mu_2^2}{\theta} + \left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}\left[2\mu_1\mu_2 + \mu_2^2 - \frac{\mu_2^2}{\theta} - \left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}\mu_2^2\right]$
SVB Poisson-NB	$E(X) + \frac{\mu^2}{\alpha}$
SVB NB-NB	$E(X) + \frac{\mu_1^2}{\alpha} + \frac{\mu_2^2}{\theta}$

Since the variances of the SVA and SVB models considered are always greater than their expectations, these models are suitable for over-dispersed data.

3.4.4 Probability generating functions of SVA and SVB distributions

If X is a random variable taking non-negative integer values, then its probability generating function (pgf) is defined as:

$$P_x(t) = E(t^X) = \sum_{x=0}^{\infty} t^x \cdot P(X = x) \tag{3.44}$$

Note that if a distribution X has mgf, $M_X(t)$ and pgf, $G_X(t)$, then

$$G_X(e^t) = M_X(t) \tag{3.45}$$

The convolution theorem in (3.12) also holds when computing the pgf. The Poisson distribution has pgf:

$$\begin{aligned} P_x(t) &= E(t^X) \\ &= \sum_{x=0}^{\infty} t^x \, P(X = x) \\ &= \sum_{x=0}^{\infty} t^x \, \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} \\ &= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda t)^x}{x!} \\ &= \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda t} \\ &= \mathrm{e}^{\lambda(t-1)} \end{aligned} \tag{3.46}$$

The negative binomial distribution has pgf:

$$\begin{aligned} P_x(t) &= E(t^X) \\ &= \sum_{x=0}^{\infty} t^x \binom{x+\alpha-1}{\alpha-1} p^\alpha (1-p)^x \\ &= p^\alpha \sum_{x=0}^{\infty} \binom{x+\alpha-1}{\alpha-1} (t(1-p))^x \\ &= p^\alpha \sum_{x=0}^{\infty} \binom{x+\alpha-1}{x} (t(1-p))^x \\ &= p^\alpha \sum_{x=0}^{\infty} (-1)^x \binom{-\alpha}{x} (t(1-p))^x \\ &= p^\alpha \sum_{x=0}^{\infty} \binom{-\alpha}{x} (t(1-p))^x \\ &= p^\alpha \, (1 - t(1-p))^{-\alpha} \\ &= \left(\frac{p}{1-t(1-p)} \right)^\alpha \end{aligned} \tag{3.47}$$

Note that the mean and variance can also be derived from the pgf:

$$E(X) = P'_x(t) \tag{3.48}$$

$$Var(X) = P''_x(1) + P'_x(1) - (P'_x(1))^2 \tag{3.49}$$

The pgfs of the proposed SVA and SVB distributions are given in Table 3.3. The expectations and variances here are similar to those given in Section 3.4.3.

Table 3.3: *Pgfs, expectations and variances of SVA and SVB distributions considered.*

Distributions	Pgf	E(X)	Var(X)
SVA PoisNB	$\mathrm{e}^{-\lambda}\left(1-\left(\frac{p}{1-t(1-p)}\right)^{\alpha}\right)+\mathrm{e}^{\lambda(t-1)}\left(\frac{p}{1-t(1-p)}\right)^{\alpha}$	$\lambda+\mu-\mathrm{e}^{-\lambda}\mu$	$E(X)+\frac{\mu^2}{\alpha}+\mathrm{e}^{-\lambda}\left(2\lambda\mu+\mu^2-\mathrm{e}^{-\lambda}\mu^2-\frac{\mu^2}{\alpha}\right)$
SVA NBPois	$p^{\alpha}\left(1-\mathrm{e}^{\lambda(t-1)}\right)+\left(\frac{p}{1-t(1-p)}\right)^{\alpha}\mathrm{e}^{\lambda(t-1)}$	$\lambda+\mu-\lambda\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}$	$E(X)+\frac{\mu^2}{\alpha}+\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}\left[\lambda^2+2\lambda\mu-\left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}\lambda^2\right]$
SVA NBNB	$p^{\alpha}\left(1-\left(\frac{q}{1-(1-q)t}\right)^{\theta}\right)+\left(\frac{p}{1-(1-p)t}\right)^{\alpha}\left(\frac{q}{1-(1-q)t}\right)^{\theta}$	$\mu_1+\mu_2-\mu_2\left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}$	$E(X)+\frac{\mu_1^2}{\alpha}+\frac{\mu_2^2}{\theta}+\left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}\left[2\mu_1\mu_2+\mu_2^2-\frac{\mu_2^2}{\theta}-\left(\frac{\alpha}{\mu_1+\alpha}\right)^{\alpha}\mu_2^2\right]$
SVB PoisNB	$\mathrm{e}^{\lambda(t-1)}\left(\frac{p}{1-t(1-p)}\right)^{\alpha}$	$\lambda+\mu$	$E(X)+\frac{\mu^2}{\alpha}$
SVB NBNB	$\left(\frac{p}{1-t(1-p)}\right)^{\alpha}\left(\frac{q}{1-t(1-q)}\right)^{\theta}$	$\mu_1+\mu_2$	$E(X)+\frac{\mu_1^2}{\alpha}+\frac{\mu_2^2}{\theta}$

3.4.5 Skewness and kurtosis of SVA and SVB distributions

The skewness and kurtosis of distributions are aspects of their shape. Skewness measures symmetry. For example, the skewness of a normal distribution is zero as it is symmetric about its mean. The coefficient of skewness is determined using the third moment:

$$\begin{aligned}\gamma_1 &= \frac{E(X-\mu)^3}{\sigma^3} \\ &= \frac{\mu_3}{\mu_2^{3/2}} \\ &= \frac{1}{(\sigma^2)^{3/2}}\left(E(X^3) - 3\mu E(X^2) + 2\mu^3\right)\end{aligned} \tag{3.50}$$

Here, μ_3 is the third central moment of X, where $\mu_3 = E(X-\mu)^3$, and the second central moment, μ_2, is the variance. The skewness of some standard distributions are:

(i) Poisson(λ)

$$\gamma_1 = \frac{1}{\sqrt{\lambda}} \tag{3.51}$$

(ii) Negative binomial, NB2(μ, α)

$$\gamma_1 = \frac{\alpha + 2\mu}{\sqrt{\alpha\mu(\mu+\alpha)}} \tag{3.52}$$

(iii) Neyman type A(λ, ϕ)

$$\gamma_1 = \frac{\phi^2 + 3\phi + 1}{(\phi+1)\sqrt{\lambda\phi\,(\phi+1)}} \tag{3.53}$$

(iv) Compound Poisson-NB(λ, μ, α)

$$\gamma_1 = \frac{\alpha\mu(\mu+3) + \mu^2(3+2\alpha^{-1}) + 3\mu + \alpha}{(\alpha\mu+\alpha+\mu)\sqrt{\mu\lambda(\mu+\mu\alpha^{-1}+1)}} \tag{3.54}$$

(v) Compound NB-Poisson(μ, α, λ)

$$\gamma_1 = \frac{\lambda^2(\alpha+3\mu+2\mu^2\alpha^{-1}) + 3\alpha\lambda + 3\mu\lambda + \alpha}{(\alpha\lambda+\mu\lambda+\alpha)\sqrt{\mu\lambda(1+\lambda+\mu\lambda\alpha^{-1})}} \tag{3.55}$$

(vi) Compound NB-NB($\mu_1, \alpha_1, \mu_2, \alpha_2$)

$$\begin{aligned}\gamma_1 =&\Bigg(\mu_2^2\Big(\alpha\theta + 3\mu_1\theta + 3\mu_1 + 3\alpha + 2\theta\mu_1^2\alpha^{-1} + 2\alpha\mu_2\theta^{-1}\Big)\\&+ 3\mu_2\Big(\alpha\theta + \theta\mu_1 + \alpha\Big) + \alpha\theta\Bigg)\cdot\Bigg(\alpha\theta\mu_2 + \theta\mu_1\mu_2 + \alpha\theta + \alpha\mu_2\Bigg)^{-1}\\&\cdot\Bigg(\mu_1\mu_2\Big(1 + \mu_2 + \mu_2\theta^{-1} + \mu_1\mu_2\alpha^{-1}\Big)\Bigg)^{1/2}\end{aligned} \tag{3.56}$$

In all cases below, let $K = \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha}$. For SVA NB-NB, $T = \left(\frac{\alpha}{\mu_1+\alpha}\right)$. The skewness of SVA and SVB distributions are:

(i) SVA Poisson-NB

$$\begin{aligned}&\Bigg(e^{-2\lambda}\alpha\mu^2\left(-2e^{-\lambda}\mu + 6\lambda + 3\mu - 3\right) + e^{-\lambda}\mu\Big(-3e^{-\lambda}\mu^2 - 3\alpha\lambda^2\\&- 3\alpha\lambda\mu - \alpha\mu^2 + 6\alpha\lambda + 3\alpha\mu + 3\lambda\mu + 3\mu^2 - \alpha - 3\mu - 2\mu^2\alpha^{-1}\Big)\\&+ \alpha\lambda + \alpha\mu + 3\mu^2 + 2\mu^3\alpha^{-1}\Bigg)\cdot\\&\Bigg(\mu e^{-\lambda}\Big(e^{-\lambda}\mu - 2\lambda - \mu + 1 + \mu\alpha^{-1}\Big) - \Big(\lambda + \mu + \mu^2\alpha^{-1}\Big)\Bigg)^{-1/2}\cdot\\&\Bigg(\lambda\alpha + \mu\alpha + \mu^2 - \alpha\mu e^{-\lambda}\Big(e^{-\lambda}\mu - 2\lambda - \mu + 1 + \alpha^{-1}\mu\Big)\Bigg)^{-1}\end{aligned} \tag{3.57}$$

(ii) SVA NB-Poisson

$$\begin{aligned}&\Bigg(K\alpha\lambda\Big(\lambda^2 + 3\lambda\mu + 3\mu^2 - 3\lambda - 6\mu + 1\Big) + K^2\alpha\lambda^2\Big(2K\lambda - 3\lambda\\&- 6\mu + 3\Big) - 3K\lambda\mu^2 - \lambda\alpha - \alpha\mu - 3\mu^2 - 2\mu^3\alpha^{-1}\Bigg)\cdot\\&\Bigg(K\lambda\Big(-K\lambda + \lambda + 2\mu - 1\Big) + \lambda + \mu + \mu^2\alpha^{-1}\Bigg)^{-1/2}\cdot\\&\Bigg(K\lambda\Big(K\alpha\lambda - \alpha\lambda - 2\alpha\mu + \alpha\Big) - \lambda\alpha - \alpha\mu - \mu^2\Bigg)\end{aligned} \tag{3.58}$$

(iii) SVA NB-NB

$$\Bigg(T^2\alpha\theta\mu_2^2\Big(2T\mu_2-6\mu_1-3\mu_2+3\Big)-T\alpha\theta\mu_2\Big(-3\mu_1^2-3\mu_1\mu_2 \\ -\mu_2^2+6\mu_1+3\mu_2-1\Big)-T\alpha\mu_2^2\Big(-3T\mu_2+3\mu_1+3\mu_2-3 \\ -2\theta^{-1}\mu_2\Big)-3T\theta\mu_1\mu_2-\alpha\theta\Big(\mu_1+\mu_2+2\theta^{-2}\mu_2^3+2\alpha^{-2}\mu_1^3\Big) \\ +3\alpha\mu_2^2+3\theta\mu_1^2\Bigg)\cdot\Bigg(T\mu_2\Big(T\mu_2-2\mu_1-\mu_2+\theta^{-1}\mu_2+1\Big) \\ -\mu_1-\mu_2-\theta\mu_2^2-\mu_1^2\alpha^{-1}\Bigg)^{-1/2}\cdot\Bigg(T\alpha\mu_2\Big(2\theta\mu_1-T\theta\mu_2+\theta\mu_2 \\ -\theta-\mu_2\Big)+\alpha\theta\Big(\mu_1+\mu_2\Big)+\alpha\mu_2^2-\theta\mu_1^2\Bigg)^{-1} \tag{3.59}$$

(iv) SVB Poisson-NB

$$\Big(\alpha\lambda+\alpha\mu+3\mu^2-2\mu^3\alpha^{-1}\Big)\cdot\Big(\alpha\lambda+\alpha\mu+\mu^2\Big)^{-1}\cdot\Big(\lambda+\mu+\mu^2\alpha^{-1}\Big)^{-1/2} \tag{3.60}$$

(v) SVB NB-NB

$$\Bigg(\alpha\theta\mu_1+\alpha\theta\mu_2+3\alpha\mu_2^2+2\alpha\mu_2^3\theta^{-1}+3\theta\mu_1^2+2\theta\mu_1^3\alpha^{-1}\Bigg)\cdot \\ \Bigg(\alpha\theta\Big(\mu_1+\mu_2\Big)+\alpha\mu_2^2+\theta\mu_1^2\Bigg)^{-1}\cdot\Bigg(\mu_1+\mu_2+\mu_2^2\theta^{-1}+\mu_1^2\alpha^{-1}\Bigg)^{-1/2} \tag{3.61}$$

Kurtosis measures the heaviness of the tail of distributions. This is determined using the fourth moment:

$$\begin{aligned} kurtosis &= \frac{E(X-\mu)^4}{\sigma^4} \\ &= \frac{\mu_4}{\mu_2^2} \\ &= \frac{E(X^4)-4\mu E(X^3)+6\mu^2E(X^2)-3\mu^4}{(\sigma^2)^2} \end{aligned} \tag{3.62}$$

The kurtosis of a normal distribution is 3 and the kurtosis of a distribution is

often compared to that of a normal distribution, where the kurtosis coefficient is:

$$\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4} - 3 \tag{3.63}$$

A positive kurtosis coefficient imply that the distribution has a heavier tail or sharper compared to the normal distribution, while a negative kurtosis coefficient imply that the distribution is lighter tailed or flatter than the normal distribution (Dodge, 2008). The kurtosis coefficients of some standard distributions are:

(i) Poisson(λ)

$$\gamma_2 = \frac{1}{\lambda} \tag{3.64}$$

(ii) Negative binomial, NB2(μ, α)

$$\gamma_2 = \frac{\alpha^2 + 6\alpha\mu + 6\mu^2}{\alpha\mu(\mu+\alpha)} \tag{3.65}$$

(iii) Neyman type A(λ, ϕ)

$$\gamma_2 = \frac{\phi^3 + 6\phi^2 + 7\phi + 1}{\lambda\phi(\phi+1)^2} \tag{3.66}$$

(iv) Compound Poisson-NB(λ, μ, α)

$$\gamma_2 = \frac{\alpha^2\mu(6+\mu+\mu^{-2}) + 6\alpha\mu(3+\mu+\mu\alpha^{-2}) + \mu(11\mu+12) + 7\alpha(\alpha+1)}{\lambda(\alpha\mu+\alpha+\mu)^2} \tag{3.67}$$

(v) Compound NB-Poisson(μ, α, λ)

$$\gamma_2 = \frac{7\alpha\lambda(\mu\lambda^2+\alpha+\mu) + \lambda^3(\alpha^2+12\mu^2+6\mu^3\alpha^{-1}) + 6\lambda^2(\alpha^2+3\alpha\mu+2\mu^2) + \alpha^2}{\mu\lambda(\alpha\lambda+\mu\lambda+\alpha)^2} \tag{3.68}$$

(vi) Compound NB-NB$(\mu_1, \alpha_1, \mu_2, \alpha_2)$

$$\begin{aligned}\gamma_2 =& \Bigg(\alpha^2\theta^2\Big(7 + \mu_2^2 + 6\mu_2 + \mu_2^{-1}\Big) + \alpha\theta^2\mu_1\Big(7 + 7\mu_2^2 + 18\mu_2 + 6\mu_1^2\mu_2^2\alpha^{-2}\Big) \\ &+ \alpha^2\theta\Big(7 + 18\mu_2 + 6\mu_2^2 + 6\mu_2^2\theta^{-2}\Big) + 18\alpha\theta\mu_1\mu_2\Big(\mu_2 + 1\Big) \\ &+ 12\theta^2\mu_1\mu_2\Big(\mu_1\mu_2 + \mu_1\Big) + \alpha\mu_2\Big(11\alpha\mu_2 + 11\mu_1\mu_2 + 12\alpha\Big) + 12\theta\mu_1^2\mu_2^2\Bigg) \\ &\cdot \left(\mu_1\Big(\alpha\theta\mu_2 + \theta\mu_1\mu_2 + \alpha\theta + \alpha\mu_2\Big)^2\right)^{-1}\end{aligned} \tag{3.69}$$

Hence these distributions are sharper than the normal distribution.

The kurtosis coefficients of the SVA and SVB distributions considered are:

(i) SVA Poisson-NB

$$\begin{aligned}&\Bigg(\mathrm{e}^{-\lambda}\mu\Big(-18\alpha\mu^2 + 7\alpha\mu + 12\mu^2 - \alpha^2\mu^3 + 6\alpha^2\mu^2 + 6\alpha\mu^3 - 7\alpha^2\mu \\ &- 11\mu^3 + \alpha^2 - 4\alpha^2\lambda^3 - 6\alpha^2\lambda^2\mu - 4\alpha^2\lambda\mu^2 + 18\alpha^2\lambda^2 + 18\alpha^2\lambda\mu \\ &+ 6\alpha\lambda^2\mu + 12\alpha\lambda\mu^2 - 14\alpha^2\lambda - 8\lambda\mu^2 - 18\alpha\lambda\mu\Big) \\ &+ \mathrm{e}^{-2\lambda}\alpha\lambda\mu\Big(24\alpha\lambda\mu + 24\alpha\mu^2 - 36\alpha\mu - 24\mu^2\Big) \\ &+ \mathrm{e}^{-2\lambda}\alpha\mu\Big(7\alpha\mu^3 - 18\alpha\mu^2 - 18\mu^3 + 7\alpha\mu + 18\mu^2\Big) + 11\mathrm{e}^{-2\lambda}\mu^4 \\ &+ \mathrm{e}^{-3\lambda}\alpha\mu\Big(-24\alpha\lambda\mu^2 + 12\mu^3 - 12\alpha\mu^3 + 12\alpha\mu^2\Big) \\ &+ 6\mu^4\Big(\mathrm{e}^{-4\lambda}\alpha^2\mu^4 + \mathrm{e}^{-\lambda} - 1\Big) + \alpha^2\lambda - \alpha^2\mu - 7\alpha\mu^2 - 12\mu^3\Bigg)\cdot \\ &\left(\mathrm{e}^{-2\,\lambda}\alpha\,\mu^2 - 2\,\mathrm{e}^{-\lambda}\alpha\,\lambda\,\mu - \mathrm{e}^{-\lambda}\alpha\,\mu^2 + \mathrm{e}^{-\lambda}\alpha\,\mu + \mathrm{e}^{-\lambda}\mu^2 - \lambda\,\alpha - \mu\,\alpha - \mu^2\right)^{-2}\end{aligned} \tag{3.70}$$

(ii) SVA NB-Poisson

$$\Bigg(K\alpha\lambda\mu\Big(24K^2\lambda^2\alpha - 24K\lambda^2\alpha - 24K\lambda\mu\alpha + 36K\lambda\alpha + 12K\lambda\mu - 6\lambda\mu - 12\mu^2 - 18\lambda\alpha - 18\mu\alpha + 4\lambda^2\alpha + 6\lambda\mu\alpha + 4\mu^2\alpha + 14\alpha + 18\mu\Big) + K\lambda\alpha\Big(12K^2\lambda^3\alpha - 12K^2\lambda^2\alpha + 18K\lambda^2\alpha - 7K\lambda\alpha - 7K\lambda^3\alpha + \lambda^3\alpha + 7\lambda\alpha - 6\lambda^2\alpha - \alpha\Big) + 6K^4\lambda^4\alpha^2 + 8K\lambda\mu^3 + \alpha^2\mu + \alpha^2\lambda + 6\mu^4\alpha^{-1} + 7\alpha\mu^2 + 12\mu^3\Bigg) \cdot \Bigg(K\alpha\lambda\Big(-K\lambda + \lambda + 2\mu - 1\Big) + \lambda\alpha + \alpha\mu + \mu^2\Bigg)^{-2} \tag{3.71}$$

(iii) SVA NB-NB

$$\begin{aligned}
&\Bigg(12T^3\mu_2^3\alpha^2\theta\Big(\mu_2 + \theta - 2\mu_1\theta - \mu_2\theta\Big) \\
&+ 12T^2\mu_1\mu_2^2\alpha^2\theta^2\Big(2\mu_1 + 2\mu_2 - \mu_1\alpha^{-1} - 2\mu_2\theta^{-1} - 3\Big) \\
&+ 6T^2\mu_2^3\alpha^2\theta\Big(T^2\mu_2\theta - 3\mu_2 - 3\theta + 3\Big) \\
&+ T^2\mu_2^2\alpha^2\theta\Big(7\mu_2^2\theta + 11\mu_2^2\theta^{-1} + 7\theta\Big) \\
&+ 6T\mu_1\mu_2\alpha^2\theta\Big(\mu_1\mu_2 - 3\mu_2 - 3\mu_1\alpha^{-1}\theta - \mu_1\mu_2\theta + 3\mu_1\theta \\
&- \mu_1\mu_2\alpha^{-1} + 3\mu_2\theta + 2\mu_2^2 + 2\mu_1^2\alpha^{-1}\theta + \mu_1\mu_2\alpha^{-1}\theta\Big) \\
&- 2T\mu_1\mu_2\alpha^2\theta^2\Big(2\mu_1^2 + 2\mu_1\mu_2^2 + 4\mu_1^2\alpha^{-2} + 4\mu_2^2\theta^{-2} + 7\Big) \\
&+ 6T\mu_2^3\alpha^2\theta\Big(\theta + \mu_2 + \mu_2\theta^{-2} + 2\theta^{-1} - 3\Big) \\
&- T\mu_2\alpha^2\theta\Big(11\mu_2^3 + 7\mu_2\theta + \mu_2^3\theta - 7\mu_2 - \theta\Big) \\
&- \alpha^2\mu_2^2\Big(6\mu_2^2\alpha^2\theta^{-1} + 12\mu_2 + 7\theta\Big) \\
&- \theta^2\mu_1^2\Big(6\mu_1^2\theta^2\alpha^{-1} + 12\mu_1 + 7\alpha\Big) - \alpha^2\theta^2\Big(\mu_1 + \mu_2\Big)\Bigg)\cdot \\
&\Bigg(T\alpha\theta\mu_2\Big(T\mu_2 - 2\mu_1 - \mu_2 + \mu_2\theta^{-1} + 1\Big) - \alpha\theta\Big(\mu_1 + \mu_2\Big) \\
&- \alpha\mu_2^2 - \theta\mu_1^2\Bigg)^{-2}
\end{aligned} \tag{3.72}$$

(iv) SVB Poisson-NB

$$\Big(\alpha^2\lambda + \alpha^2\mu + 7\alpha\mu^2 + 12\mu^3 + 6\mu^4\alpha^{-1}\Big) \cdot \Big(\alpha\lambda + \alpha\mu + \mu^2\Big)^{-2} \tag{3.73}$$

(v) SVB NB-NB

$$\begin{aligned}\Bigg(&\alpha^2\theta^2\Big(\mu_1 + \mu_2\Big) + \alpha^2\mu_2^2\Big(7\theta + 12\mu_2 + 6\mu_2\theta^{-1}\Big) \\ &+ \theta^2\mu_1^2\Big(7\alpha + 12\mu_1 + 6\mu_1^2\alpha^{-1}\Big)\Bigg) \cdot \Bigg(\alpha\theta\Big(\mu_1 + \mu_2\Big) + \alpha\mu_2^2 + \theta\mu_1^2\Bigg)^{-2}\end{aligned} \tag{3.74}$$

3.5 Model fitting algorithms

In this section, we discuss in detail how the pmf of the standard compound models and their variants are computed in R (R Core Team, 2014).

3.5.1 Computation of standard compound pmf

Recall that in a standard compound distribution, the case $X = 0$ implies that the second generation is zero, regardless of the values in the first generation. It is necessary to take into account the number of zero(s) which are present in the second generation when determining the probabilities. Hence, given X which follows a compound A-B distribution, where the first and second generations follow distributions with pmf $a(x)$ and $b(x)$ respectively, then in general:

$$P(X = 0) = a(0) + \sum_{i=1}^{\infty} a(i)\,(b(0))^i \tag{3.75}$$

Suppose that compound models are labelled as in Table 3.4, then the computation of $P(X = k)$, where $k > 0$, can be obtained based on the reasoning in Table 3.5.

Table 3.4: *Generations in compound model*

First generation	i
Second generation	k
Total	$k + i$

In Table 3.5, each row in the second generation sums to k and the number of

Table 3.5: *Possible combinations in compound models*

For $X = 1$:

First gen.	1	2	3	4	...
Second gen.	1	1 0	1 0 0	1 0 0 0	...

For $X = 2$:

First gen.	1	2	3	4	...
Second gen.	2	2 0 1 1	2 0 0 1 1 0	2 0 0 0 1 1 0 0	...

For $X = 3$:

First gen.	1	2	3	4	...
Second gen.	3	3 0 2 1	3 0 0 2 1 0 1 1 1	3 0 0 0 2 1 0 0 1 1 1 0	...

etc.

entries is based on the corresponding value in the first generation. The unique entries in the second generation can be obtained using the *restrictedparts* command in the package *partitions* (Hankin, 2006) in R. It is also necessary to consider the number of distinguishable permutations when determining the probabilities for each row in the second generation in Table 3.5. This is done by dividing the factorial of the total number of entries for each row, by the frequency of each number in that entry:

$$\frac{N!}{(n_1!)(n_2!)(n_3!)\dots(n_k!)} \tag{3.76}$$

For example, for the row containing (*2 1 0 0*) in $P(X = 3)$, the distinguishable permutation is $\frac{4!}{1!1!2!}$. Hence, using $X = 3$ as an example, the individual probability for each i^{th} entry in the first generation is in Figure 3.1, and the sum of all these probabilities will yield $P(X = 3)$. These calculations can be performed in R and an example will be discussed in the next section.

Neyman type A pmf

The R code in Listing 3.1 incorporates the steps of Figure 3.1 to compute the pmf of the Neyman type A distribution. Some line comments are given after the # symbol. Since the Neyman type A is a compound Poisson-Poisson, both A and B are Poisson distributions, with parameters `lam1` and `lam2` respectively.

Figure 3.1: *Computation for P(X=3) in compound model*

i	$P(X=3\|i)$
1	$a(1) \times b(3)$
2	$a(2) \times (b(3)b(0) \times 2! + b(2)b(1) \times 2!)$
3	$a(3) \times \left(b(3)(b(0))^2 \times \frac{3!}{2!} + b(2)b(1)b(0) \times 3! + (b(1))^3 \times \frac{3!}{3!}\right)$
4	$a(4) \times \left(b(3)(b(0))^3 \times \frac{4!}{3!} + b(2)b(1)(b(0))^2 \times \frac{4!}{2!} + (b(1))^3 b(0) \times \frac{4!}{3!}\right)$
...	...
Total	$P(X=3) = \sum_{i=1}^{\infty} P(X=3\|i)$

```
1  NeyAfun = function(lam1, lam2, x){
2    if (x==0) {dpois(0, lambda=exp(lam1)) + sum(dpois(1:500, lambda=
         exp(lam1))*dpois(0,lambda=exp(lam2))^(1:500))} else
3      if (x==1) {sum(dpois(1:500, lambda=exp(lam1))*dpois(1, lambda=
           exp(lam2))*dpois(0,lambda=exp(lam2))^(0:499)*(1:500) ) }
4    else{
5      yy=list(matrix())
6      Prob=numeric()
7      Prob2=numeric()
8      for(i in 1:50){ #we want to do this from 1 to infty
9        yy[[i]] = restrictedparts(x,i+1) #use i+1 so that it always
             starts with 2 partitions
10       dim2 = function(s) dim(s)[2]
11       dd = unlist(lapply(yy,dim2)) #number of columns in each matrix
12       for(j in 1:dd[i]){
13         yyj = yy[[i]][,j] #jth column of the matrix
14         Uniq_entries = unique(yyj) #list of unique entries in the
               column
15         No_entries_fun =function(r) length(yyj[yyj==r]) #function to
               calculate number of times each unique entry in the
               column occurs
16         No_entries = mapply(No_entries_fun,Uniq_entries) #Number of
               times each entry occur
17         Combi = factorial(length(yyj))/prod(factorial(mapply(No_
               entries_fun,Uniq_entries))) #number of ways to get each
               combo of jth column in the matrix
18         Prob2[j] = prod(dpois(Uniq_entries,lambda=exp(lam2))^(No_
               entries)) * Combi
19       }
20       Prob[i] = dpois(i+1,lambda=exp(lam1))*sum(Prob2)
21     }
22     return(sum(c(dpois(1,lambda=exp(lam1))*dpois(x,lambda=exp(lam2))
           ,Prob)))
23   }
24 }
```

Listing 3.1: *R code to calculate the probabilities for the Neyman type A model*

The code in Listing 3.1 is separated into three parts, lines 2 and 3 compute $P(X = 0)$ and $P(X = 1)$ respectively, whilst the following lines compute $P(X \geq 2)$. Although theoretically this should be computed "up to infinity", this is only done up to 500 for $X < 2$, and up to 50 for $X \geq 2$, as these should be sufficient for small values of *lam1* and *lam2*; this also reduces the computation time. Note that the `dpois` command returns the Poisson probabilities based on the given parameter.

Line 5 creates a list of empty matrices to accommodate all the possible partitions (for each given x value) as of the examples in Table 3.5. Line 9 fills in the list with each of the possible combinations as a matrix. For example, if $X = 3$, then $yy[[1]]$ is a 2×2 matrix, $yy[[2]]$ is a 3×3 matrix and so on. *i+1* is used in line 9 as *restrictedparts* only splits numbers to two or more partitions. Line 10 creates a function *dim2* to obtain the number of columns, and line 11 gives *dd*, which is a vector consisting of number of columns in each matrix in the list.

Line 16 computes the number of times each unique entry in the matrix column occurs. This is then used in line 17 which is equivalent to (3.76). Note that the `prod` command returns the product of the given values.

Line 18 gives the probabilities of the second generation (right hand side of Figure 3.1) while line 19 combines it with $a(i)$ for $i \geq 2$. Finally, line 23 sums all the probabilities to return $P(X = x)$.

Oliveira et al. (2016) proposed a different function to fit the Neyman type A, where a recurrence relation is used. However, the code used here can be adapted to fit any compound model using slight modifications, that is, by changing the `dpois` command for the Poisson distribution to a function from the distribution of interest for the first and/or second generation. Dobbie and Welsh (2001) also highlighted that model fitting for the Neyman type A in particular is complicated by infinite sums in its pmf.

3.5.2 Computation of SVA and SVB pmf

The computation of SVA models is more straight forward. For example, in the case when there are no covariates, the pmf of the SVA Poisson-NB model is obtained by running the code given in Listing 3.2, where $a0$ is the Poisson parameter, and $b0$ and $c0$ are the negative binomial parameters. Note that in all cases, a log-link is used to fit the models.

```
1 svaPoisNBfun = function(a0,b0,c0, x) if (x==0) {exp(-exp(a0))} else
    {sum(dpois(1:x, exp(a0))*dnbinom((x-(1:x)), size = exp(c0) , mu =
     exp(b0)))}
2 svaPoisNBden = function(a0,b0,c0, x) mapply(svaPoisNBfun,a0,b0,c0, x
    )
```

Listing 3.2: *R code to calculate the probabilities for the SVA Poisson-NB model*

For SVB models, the 'if' function is not needed as SVB models do not follow the zero restriction. Sample code for the SVB Poisson-NB model is given in Listing 3.3.

```
1 svbPoisNBfun = function(a0,b0,c0,x) sum(dpois(0:x, exp(a0))*dnbinom
    ((x-(0:x)), size = exp(c0) , mu = exp(b0)))
2 svbPoisNBden = function(a0,b0,c0, x) mapply(svbPoisNBfun,a0,b0,c0, x
    )
```

Listing 3.3: *R code to calculate the probabilities for the SVB Poisson-NB model*

3.6 Methods of parameter estimation

3.6.1 Optimisation processes

Although previous research has emphasized the importance of having good initial parameter estimates, it is also recognised that this is not always straightforward (Dobbie and Welsh, 2001). In the case of zero-inflated models for biodosimetry data, Oliveira et al. (2016) used the Poisson estimates as a starting point for the estimation of mean of their various models, and the estimates obtained from logistic regression for the zero parts of the zero-inflated models. Dobbie and Welsh (2001) suggest that the Newton-Raphson method will work reasonably well if the parameters are roughly uncorrelated. We use the general purpose optimisation function (*optim*) in R to obtain parameters which minimise the respective functions used. Although the minimisation procedure is the default in *optim* and we wish to obtain the parameter estimates which maximise the log-likelihood function of our models, this can be achieved by minimising the negative of the log-likelihood function. Also, the *optim* function in R allows a maximisation problem if *control$fnscale* is set to negative.

3.6.2 EM algorithm

The expectation maximisation (EM) algorithm is a two step iterative process which was first introduced by Ceppellini et al. (1955). The EM algorithm is commonly used to estimate parameters when there are missing data (Dempster

et al., 1977; Do and Batzoglou, 2008). As the name suggest, this algorithm is a two step iterative process to estimate parameters which will give a maximum likelihood value.

Given an observed data set Y, and log-likelihood function $f(X|\phi)$ for the full data set, we wish to estimate the set of unknown parameters θ, by maximising

$$Q(\phi|\phi^n) = E(\log f(X|\phi)|Y, \phi^n) \tag{3.77}$$

where ϕ^n is the estimate for the n^{th} iteration.

(i) **E-step**: Using some initial parameter estimates (ϕ^n), compute the conditional expectation, $Q(\phi|\phi^n)$.

(ii) **M-step**: Choose ϕ^{n+1} which maximises $Q(\phi|\phi^n)$.

(iii) Repeat these steps until a convergence is achieved. (Becker et al., 1997)

Although the EM algorithm is easier to program, Lambert (1992) found that the Newton-Raphson algorithm is faster when maximising log-likelihood for zero-inflated Poisson models.

The EM algorithm was considered to estimate parameters whilst simulating standard compound models. Nonetheless, the EM algorithm did not increase the computation speed and hence was not used in further computation.

Chapter 4

Simulation studies

4.1 Introduction

In this chapter, we present some simulation studies to assess the fits of standard compound models and their variants using model selection criterion. In Section 4.2, we generated data using standard compound models and fitted the data using Poisson models, standard negative binomial models, standard compound models and variants of compound models. In Section 4.3, data is simulated using SVA and SVB distributions. We discover some unique properties of the variant models using probability plots, which include the SVA distributions having a large $P(X = 0)$. Hence, in Section 4.4, we compare the SVA models to hurdle and zero-inflated models.

4.2 Simulation of standard compound models

In this section, we simulated data using standard compound models, with small integer parameter values, of 1 to 3 in each case, as this will result in smaller data points which will keep computation time at a minimum. For each case, 5,000 data points are simulated and a sample of 100 data points is used. The sampled data is fitted using Poisson, negative binomial, standard compound and variants of compound models. Their log-likelihood, AIC and BIC values are then recorded. This process is repeated only 25 times to limit computation time. Since computing $P(X = x)$ for the standard compound models are slow, especially as x increases (see Section 3.5.1), in the sampled data, only values of x when $x < 13$ are considered to further reduce computation time. Table 4.1 shows the average time taken to fit some simulated negative binomial data using models investigated. This was carried out using a Windows 7, 64 bit desktop with Intel® Core™ i7 processor and 8 GB RAM. It is clear that the standard compound models are

more time consuming to fit than the other models.

Table 4.1: *Computation time taken to fit models using simulated negative binomial data. The reported time is an average from* 10 *repetitions.*

Models	Average time (s)
Poisson	0.004
Negative binomial	0.013
SVA Poisson-NB	0.937
SVA NB-Poisson	1.565
SVA NB-NB	2.469
SVB Poisson-NB	1.262
SVB NB-NB	2.319
Neyman type A	838.272
Standard compound Poisson-NB	3,102.870
Standard compound NB-Poisson	2,638.072
Standard compound NB-NB	3,346.793

In the simulation studies, the model which is favoured most (out of the 25 repetitions) is summarised in Table 4.2 based on the mean value of the first generation distribution of the standard compound model.

Table 4.2: *Models selected by AIC and BIC for simulated data sets from standard compound distributions, each with* 25 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 25 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations.*

Model used to simulate data	Parameter value	Log-likelihood	AIC	BIC
Neyman type A(λ, ϕ) (each λ is paired with varying values of ϕ, where $\phi = \{1, 2, 3\}$, giving 3 cases for each λ)	$\lambda = 1$	SVA NB-NB (2/3); Neyman type A (1/3)	Neyman type A (2/3); SVA NB-NB (1/3)	Neyman type A (3/3)
	$\lambda = 2$	Neyman type A (2/3); SVA NB-NB (1/3)	Neyman type A (2/3); SVA NB-NB 1/3	Neyman type A (2/3); SVA NB-NB (1/3)
	$\lambda = 3$	Neyman type A (1/3); SVA NB-NB (2/3)	Neyman type A (2/3); SVA NB-NB (1/3)	Neyman type A (2/3); SVA NB-NB (1/3)
Standard compound Pois-NB(λ, μ, α) (each λ is paired with varying values of μ and α, where $\mu = \{1, 2, 3\}$ and $\alpha = \{1, 2, 3\}$, giving 9 cases for each λ)	$\lambda = 1$	compound NB-NB (3/9); compound Pois-NB (2/9); SVA NB-NB (2/9); NB (1/9); compound NB-Pois (1/9)	Neyman type A (5/9); compound Pois-NB (2/9); NB (2/9)	Neyman type A (6/9); NB (3/9)
	$\lambda = 2$	compound Pois-NB (6/9); NB (1/9); Neyman type A (1/9); SVA NB-NB (1/9)	Neyman type A (7/9); compound Pois-NB (1/9); NB (1/9)	Neyman type A (7/9); NB (2/9)
	$\lambda = 3$	Neyman type A (3/9); compound Pois-NB (3/9); SVA NB-NB (2/9); NB (1/9)	Neyman type A (8/9); NB (1/9)	Neyman type A (8/9); NB (1/9)

Table 4.2: *Models selected by AIC and BIC for simulated data sets from standard compound distributions, each with* 25 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 25 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations (continued).*

Model used to simulate data	Parameter value	Log-likelihood	AIC	BIC
Standard compound NB-Pois(μ, α, λ) (each μ is paired with varying values of α and λ, where $\alpha = \{1, 2, 3\}$ and $\lambda = \{1, 2, 3\}$, giving 9 cases for each μ)	$\mu = 1$	SVA NB-NB (5/9); compound NB-NB (2/9); compound NB-Pois (1/9); NB (1/9)	Neyman type A (6/9); compound NB-Pois (1/9); SVA NB-NB (1/9); NB (1/9)	Neyman type A (6/9); compound NB-Pois (2/9); NB (1/9)
	$\mu = 2$	SVA NB-NB (6/9); compound Pois-NB (3/9)	Neyman type A (9/9)	Neyman type A (9/9)
	$\mu = 3$	compound Pois-NB (5/9); SVA NB-NB (4/9)	Neyman type A (8/9); SVA Pois-NB (1/9)	Neyman type A (9/9)

Table 4.2: *Models selected by AIC and BIC for simulated data sets from standard compound distributions, each with* 25 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 25 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations (continued).*

Model used to simulate data	Parameter value	Log-likelihood	AIC	BIC
Standard compound NB-NB($\mu_1, \alpha_1, \mu_2, \alpha_2$) (each μ_1 is paired with varying values of α_1, μ_2 and α_2, where $\alpha_1 = \{1, 2, 3\}$, $\mu_2 = \{1, 2, 3\}$ and $\alpha_2 = \{1, 2, 3\}$, giving 27 cases for each μ_1)	$\mu_1 = 1$	compound NB-NB (21/27); SVA NB-NB (4/27); NB (2/27)	NB (10/27); compound NB-Pois (11/27); compound Pois-NB (3/27); SVA NB-NB (2/27); Neyman type A (1/27)	NB (17/27); compound NB-Pois (5/27); Neyman type A (4/27); compound Pois-NB (1/27)
	$\mu_1 = 2$	compound Pois-NB (16/27); SVA NB-NB (8/27); compound NB-NB (3/27)	Neyman type A (18/27); compound Pois-NB (6/27); NB (3/27)	Neyman type A (22/27); NB (4/27); compound Pois-NB/ compound NB-Pois (1/27)
	$\mu_1 = 3$	compound Pois-NB (19/27); SVA NB-NB (6/27); compound NB-Pois (2/27)	Neyman type A (23/27); NB (2/27); compound Pois-NB (1/27); compound NB-Pois (1/27)	Neyman type A (24/27); NB (3/27)

Table 4.2 shows that for simulated Neyman type A data, the SVA NB-NB model is favoured by the log-likelihood method when $\lambda = 1$ and $\lambda = 3$, but the generating model is preferred by the AIC and BIC methods, as the extra parameters in the SVA NB-NB model are penalised by these criteria.

For simulated standard compound Poisson-NB data, when $\lambda = 1$, the generating model is only preferred 2/9 times by the log-likelihood and AIC. Although the Neyman type A is not favoured by the log-likelihood, the extra parameters in the other models are penalised by both the AIC and BIC, hence the Neyman type A is preferred a majority of the time. Similar observations are obtained when $\lambda = 2$ and $\lambda = 3$. Figure 4.1 shows that both distributions are very similar for some parameter values, and hence it is possible for the Neyman type A model to be preferred by the AIC and BIC criterion even if it is not the generating model as it also has less parameters.

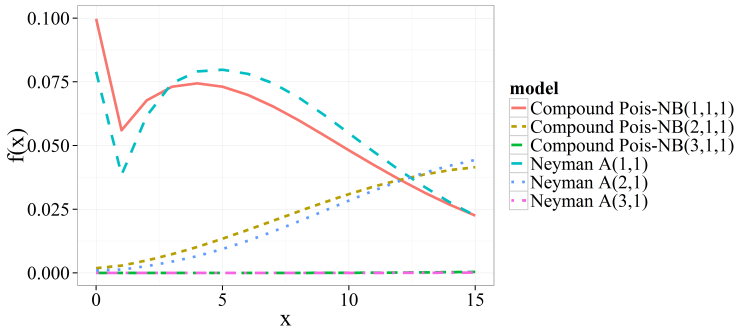

Figure 4.1: *Probability plots for Neyman type A and standard compound Poisson-NB distributions, showing that some are similar to each other for specific parameter values.*

For simulated standard compound NB-Poisson data, when $\mu = 1$, the generating model is favoured in 1/9 and 2/9 cases by AIC and BIC respectively, otherwise, the Neyman type A model is favoured most of the time. This may be due to the similarities of the two distributions for some parameter values. Given that for a fixed NB mean but larger NB size value, the NB distribution tends to a Poisson distribution, the first NB generation in the compound NB-Poisson distribution may be close to a Poisson distribution, hence it may be very similar to the Neyman type A distribution (see Figure 4.2).

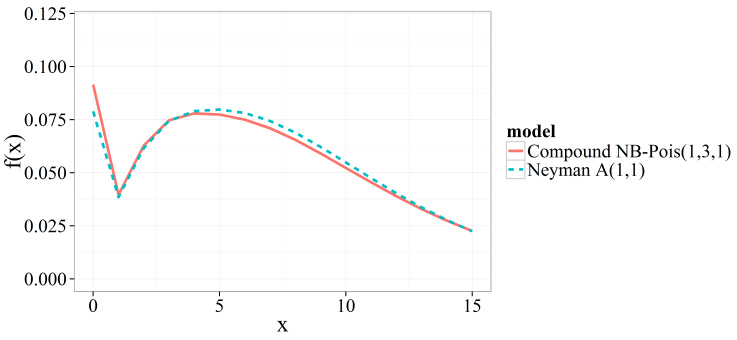

Figure 4.2: *Probability plots for Neyman type A and standard compound NB-Poisson distributions, showing that they are similar for specific parameter values.*

For simulated compound NB-NB data, the generating model is preferred by the log-likelihood in 21/27 cases when $\mu_1 = 1$, however, the extra parameters are again penalised by AIC and BIC. The compound NB-Poisson and standard negative binomial models are preferred by AIC, but the latter is preferred by BIC in most cases. When $\mu_1 = 2$, although the compound Poisson-NB model is mostly preferred by the log-likelihood method, the Neyman type A model is preferred by both AIC and BIC. Similar results also occur when $\mu_1 = 3$. This scenario is illustrated in Figure 4.3. Since the two distributions are very similar for some parameter values and the extra parameters in the standard compound Poisson-NB are penalised, the simpler Neyman type A model is favoured by the model selection criteria.

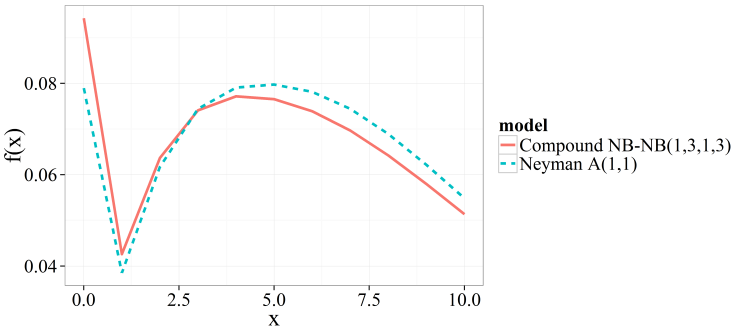

Figure 4.3: *Probability plots for Neyman type A and standard compound NB-NB distributions, showing their similarities for specific parameter values.*

4.3 Simulated data from variants of compound distributions and their preferred models

In this section, we investigate, using maximum log-likelihoods, AIC and BIC, whether the generating model will be selected when data is generated from a variant of compound distribution and refitted using the variant models. For

each of the SVA and SVB models, we generated data using varying parameter values. For example, for the simulation of SVA Poisson-NB data, we began with $\lambda = 1, \mu = 1$ and $\alpha = 1$, and then increase these by 1, up to 3, resulting in 27 different data sets. We used samples of size 1000, and refitted them to the Poisson, NB, SVA and SVB models. Since the SVA/SVB NB NB models have four parameters, 81 data sets are examined for these models. This procedure is repeated 100 times, and the proportion of times when each model is chosen, using AIC and BIC, is determined.

4.3.1 Simulation studies using SVA data

Table 4.3: *Models selected by AIC and BIC for simulated SVA data sets, each with* 100 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 100 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations.*

Model used to simulate data	parameter values	AIC	BIC
SVA Pois-NB(λ, μ, α) (each λ is paired with varying values of μ and α, where $\mu = \{1,2,3\}$ and $\alpha = \{1,2,3\}$)	$\lambda = 1$	SVA Pois-NB (8/9); NB (1/9)	SVA Pois-NB (8/9); NB (1/9)
	$\lambda = 2$	SVA NB-NB (7/9); NB (2/9)	SVA NB-NB (6/9); NB (3/9)
	$\lambda = 3$	SVA NB-NB (7/9); NB (2/9)	SVA NB-NB (4/9); NB (5/9)
SVA NB-Pois(μ, α, λ) (each μ is paired with varying values of α and λ, where $\alpha = \{1,2,3\}$ and $\lambda = \{1,2,3\}$)	$\mu = 1$	SVA NB-NB (9/9)	SVA NB-NB (9/9)
	$\mu = 2$	SVA Pois-NB (8/9); SVA NB-NB (1/9)	SVA Pois-NB (9/9)
	$\mu = 3$	SVA Pois-NB (6/9); SVA NB-NB (3/9)	SVA Pois-NB (7/9); SVA NB-NB (2/9)
SVA NB-NB($\mu_1, \alpha_1, \mu_2, \alpha_2$) (each μ_1 is paired with varying values of α_1, μ_2 and α_2, where $\alpha_1 = \{1,2,3\}$, $\mu_2 = \{1,2,3\}$ and $\alpha_2 = \{1,2,3\}$, giving 27 cases for each μ_1)	$\mu_1 = 1$	SVA NB-NB (24/27); NB (3/27)	SVA NB-NB (20/27); NB (6/27); SVA Pois-NB (1/27)
	$\mu_1 = 2$	SVA Pois-NB (20/27); SVA NB-NB (4/27); NB (3/27)	SVA Pois-NB (22/27); SVA NB-NB (2/27); NB (3/27)
	$\mu_1 = 3$	SVA Pois-NB (15/27); SVA NB-NB (8/27); NB (4/27)	SVA Pois-NB (16/27); SVA NB-NB (6/27); NB (5/27)

For the simulated SVA Poisson-NB data, when $\lambda = 1$, the generating model is selected most of the time. However, when $\lambda = 2$ and 3, SVA NB-NB is chosen as

the superior model. These results are supported by Figure 4.4, as the probability functions of the two models are very similar for some parameter values.

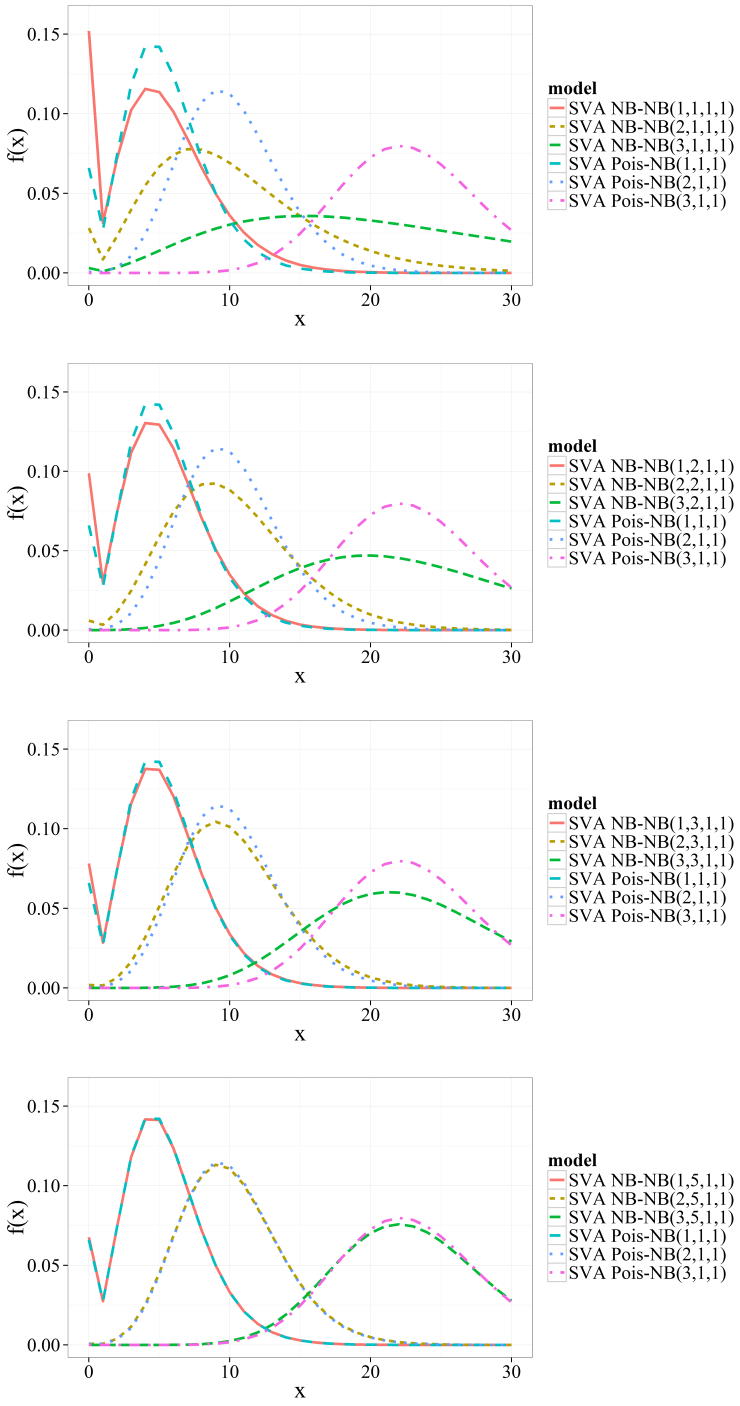

Figure 4.4: *Probability plots for SVA Poisson-NB and SVA NB-NB distributions, showing that some are similar to each other for specific parameter values.*

From Figure 4.4, if all parameters in both models are fixed, with the exception of the size parameter, α_1, of the first NB generation in SVA NB-NB in each plot, then as α_1 increases, this first NB generation converges to a Poisson distribution as the variance of NB, $\mu + \mu^2/\alpha$, converges to μ.

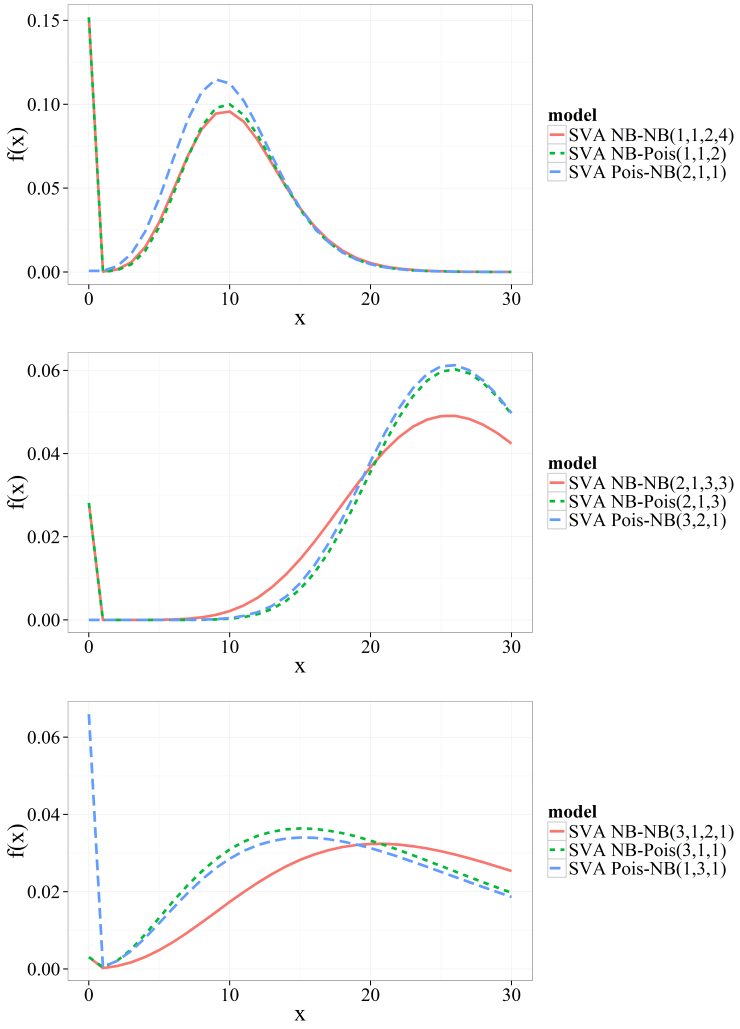

Figure 4.5: *Probability plots for SVA Poisson-NB and SVA NB-Poisson distributions, showing that some are similar to each other for specific parameter values.*

When data is simulated using the SVA NB-Poisson distribution, the SVA NB-NB model is favoured when $\mu = 1$, but the SVA Poisson-NB is favoured when $\mu = 2$ and $\mu = 3$. Some of these observations are illustrated in Figure 4.5. In these plots, the NB and Poisson generations in SVA Poisson-NB and SVA NB-Poisson models are set to have equal parameter values. In addition, the first NB generation in SVA NB-NB is also set to have equivalent parameter values with the NB generations in SVA Poisson-NB and SVA NB-Poisson. The top, middle and bottom plots represent the cases when $\mu = 1, 2, 3$ respectively. Note that the assumptions where one distribution is similar to the other by setting their parameter values are only reasonable in the absence of covariates. From Figure 4.5, it is clear that when $\mu = 1$, the SVA NB-NB distribution matches SVA NB-Poisson more closely, but the SVA Poisson-NB distribution is closer to the SVA NB-Poisson distribution when $\mu = 2$ and $\mu = 3$.

For simulated SVA NB-NB data, the generating model is favoured when $\mu = 1$,

otherwise, the SVA Poisson-NB is more likely to be chosen as the extra parameters in SVA NB-NB are penalised. Figure 4.6 also shows that the SVB NB-NB and NB distributions have very similar patterns for some parameter values. Thus, the NB model is favoured in some cases as it has fewer parameters than the SVA NB-NB model.

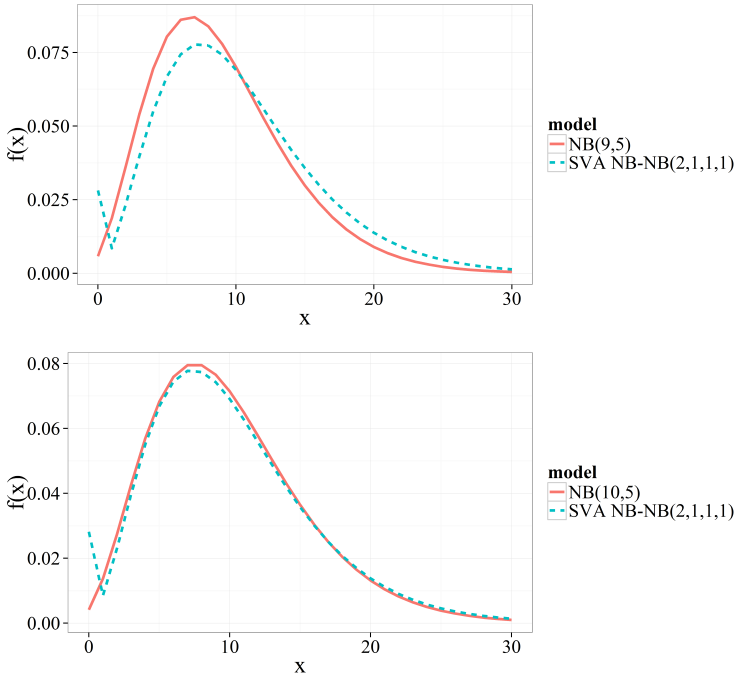

Figure 4.6: *Probability plots for SVA NB-NB and NB distributions, showing that they are similar for specific parameter values.*

4.3.2 Simulation studies using SVB data

The simulation results for simulated SVB data are given in Table 4.4. Recall that the SVB distributions are more flexible in general and resemble the negative binomial distribution to some extent. For example, Figure 4.7 shows that the SVB and NB distributions are very similar for the given parameter values.

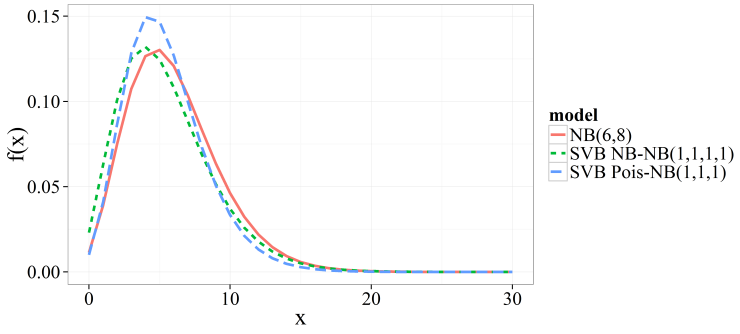

Figure 4.7: *Probability plots for SVB and NB distributions, showing that they are similar for specific parameter values.*

Table 4.4: *Models selected by AIC and BIC for simulated SVB data sets, each with* 100 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 100 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations.*

Model used to simulate data	Parameter value	AIC	BIC
SVB Pois-NB(λ, μ, α) (each λ is paired with varying values of μ and α, where $\mu = \{1, 2, 3\}$ and $\alpha = \{1, 2, 3\}$)	$\lambda = 1$	NB (5/9); SVB Pois-NB (3/9); Poisson (1/9)	NB (9/9)
	$\lambda = 2$	NB (6/9); SVB Pois-NB (2/9); Poisson (1/9)	NB (7/9); SVB Pois-NB (2/9)
	$\lambda = 3$	NB (4/9); SVB Pois-NB (3/9); Poisson (2/9)	NB (7/9); SVB Pois-NB (2/9)
SVB NB-NB($\mu_1, \alpha_1, \mu_2, \alpha_2$) (each μ_1 is paired with varying values of α_1, μ_2 and α_2, where $\alpha_1 = \{1, 2, 3\}$, $\mu_2 = \{1, 2, 3\}$ and $\alpha_2 = \{1, 2, 3\}$)	$\mu_1 = 1$	NB (23/27); SVB Pois-NB (4/27)	NB (27/27)
	$\mu_1 = 2$	NB (21/27); SVB Pois-NB (6/27)	NB (26/27); SVB Pois-NB (1/27)
	$\mu_1 = 3$	NB (21/27); SVB Pois-NB (6/27)	NB (25/27); SVB Pois-NB (2/27)

Table 4.4 shows that when data is generated using the SVB Poisson-NB distribution, the NB model is favoured most of the time. This may be due to its resemblance to the standard NB model and the penalties on the extra parameters in SVB Poisson-NB models by AIC/BIC. Similarly, for simulated SVB NB-NB data, the NB model is favoured most of the time, followed by the SVB Poisson-NB.

These results show that it is common that the generating model may not be preferred by the model selection criteria used, and hence conclusions about the most appropriate model for a data set should be made based on careful interpretation and justification.

4.4 Comparison of SVA, hurdle and zero-inflated models

Given that one of the main features of the SVA distributions is their ability to model a large number of zeros without a zero-inflation parameter, in this section, we simulated some SVA data and assessed their fit using SVA models, Poisson, negative binomial hurdle models, ZIP and ZINB models.

In the absence of covariates, the hurdle and zero-inflated distributions are different parameterisations of the same distribution (Wilson, 2008). Hence, in the absence of covariates, the ZIP log-likelihood matches that of the Poisson hurdle, and the ZINB log-likelihood matches that of a negative binomial hurdle model.

Simulations are carried out by:

(i) Simulating $5,000$ data points from a SVA distribution with no covariates.

(ii) Sampling $1,000$ data points and fit these to the Poisson hurdle, NB hurdle, ZIP, ZINB and all SVA models.

(iii) Recording and comparing their log-likelihood, AIC and BIC.

(iv) Repeating steps (ii) and (iii) to obtain 100 repetitions for each SVA distribution.

The results are presented in Table 4.5.

Table 4.5: *Models selected by log-likelihood, AIC and BIC for simulated data sets from the SVA distributions, each with* 100 *repetitions. For each combination of parameter values, the model that is mostly selected out of the* 100 *repetitions are recorded and the number in parentheses indicates this proportion out of the possible combinations.*

Model used to simulate data	Parameter value	Log-likelihood	AIC	BIC
SVA Pois-NB(λ, μ, α) (each λ is paired with varying values of μ and α, where $\mu = \{1, 2, 3\}$ and $\alpha = \{1, 2, 3\}$)	$\lambda = 1$	NB hurdle/ ZINB (8/9); SVA Pois-NB (1/9)	NB hurdle/ ZINB (6/9); SVA Pois-NB (3/9)	NB hurdle/ ZINB (6/9); SVA Pois-NB (3/9)
	$\lambda = 2$	NB hurdle/ ZINB (9/9)	NB hurdle/ ZINB (9/9)	NB hurdle/ ZINB (9/9)
	$\lambda = 3$	NB hurdle/ ZINB (9/9)	NB hurdle/ ZINB (9/9)	NB hurdle/ ZINB (9/9)
SVA NB-Pois(μ, α, λ) (each μ is paired with varying values of α and λ, where $\alpha = \{1, 2, 3\}$ and $\lambda = \{1, 2, 3\}$)	$\mu = 1$	SVA NB-Pois (6/9); SVA Pois-NB (2/9); NB hurdle/ ZINB (1/9)	SVA NB-Pois (5/9); SVA Pois-NB (3/9); NB hurdle/ ZINB (1/9)	SVA NB-Pois (5/9); SVA Pois-NB (3/9); NB hurdle/ ZINB (1/9)
	$\mu = 2$	SVA Pois-NB (5/9); SVA NB-NB (2/9); NB hurdle/ ZINB (2/9)	SVA Pois-NB (7/9); NB hurdle/ ZINB (2/9)	SVA Pois-NB (7/9); NB hurdle/ ZINB (2/9)
	$\mu = 3$	SVA Pois-NB (5/9); NB hurdle/ ZINB (3/9); SVA NB-Pois (1/9)	SVA Pois-NB (6/9); NB hurdle/ ZINB (3/9)	SVA Pois-NB (6/9); NB hurdle/ ZINB (3/9)
SVA NB-NB($\mu_1, \alpha_1, \mu_2, \alpha_2$) (each μ_1 is paired with varying values of α_1, μ_2 and α_2, where $\alpha_1 = \{1, 2, 3\}$, $\mu_2 = \{1, 2, 3\}$ and $\alpha_2 = \{1, 2, 3\}$)	$\mu_1 = 1$	NB hurdle/ ZINB (14/27); SVA Pois-NB (8/27); SVA NB-NB (5/27)	NB hurdle/ ZINB (14/27); SVA Pois-NB (13/27)	NB hurdle/ ZINB (14/27); SVA Pois-NB (13/27)
	$\mu_1 = 2$	NB hurdle/ ZINB (24/27); SVA NB-NB (2/27)	NB hurdle/ ZINB (23/27); SVA Pois-NB (4/27)	NB hurdle/ ZINB (22/27); SVA Pois-NB (5/27)
	$\mu_1 = 3$	NB hurdle/ ZINB (23/27); SVA NB-NB (3/27); SVA Pois-NB (2/27)	NB hurdle/ ZINB (25/27); SVA Pois-NB (2/27)	NB hurdle/ ZINB (25/27); SVA Pois-NB (2/27)

From Table 4.5, for simulated SVA Poisson-NB data, the NB hurdle/ZINB models are selected most of the time when $\lambda = 1$, otherwise, the generating model is selected. When $\lambda = 1$ and $\lambda = 2$, the NB hurdle/ZINB models are selected all of the time.

For simulated SVA NB-Poisson data, when $\mu = 1$, the generating model is favoured most of the time. However, when $\mu = 2$ or $\mu = 3$, SVA Poisson-NB is preferred most of the time. Again, this can be explained using Figure 4.5, as both models are very similar for some parameter values.

When data is simulated using the SVA NB-NB distribution, the NB hurdle/ ZINB models are preferred more than half of the time in all cases. Although SVA NB-NB is preferred on some occasions by the log-likelihood when $\mu_1 = 1, 2$ and 3, the extra parameters in SVA NB-NB are penalised by AIC and BIC. Hence, the NB hurdle/ ZINB models are favoured instead.

4.5 Summary

In this chapter, we showed using simulation studies that the generating model may not be preferred by AIC or BIC criterion all the time, especially in the presence of simpler models, as the extra parameters are penalised. In the absence of covariates, the probability plots show that some SVA distributions or compound distributions may be similar to each other for some specific parameter values. When SVA models are compared to hurdle and zero-inflated models, mixed results are observed as the generating models are selected on some occasions, otherwise, the other SVA models or the NB hurdle/ ZINB models are selected instead.

Chapter 5

Applications

5.1 Introduction

In this chapter, variants of compound models are applied to three real data sets. In Section 5.2, the relationship of citation counts for two time periods are investigated. In Section 5.3, compound models and their variants are applied to citation data without incorporating covariates. In Section 5.4, two relevant covariates are added, using another citation data set. Given that the variants of compound models are suitable for modelling data with a large number of zeros, Section 5.5 assesses their suitability using biodosimetry data (Oliveira et al., 2016), which involves the counts of irregular chromosomes upon exposure to radiation. Part of this chapter has been presented in conferences, published in, or submitted to journals (see Appendix A).

5.2 Citation counts as two generations

Given that it is possible to interpret citation counts as the sum of two generations, it is useful to investigate the relationship of the two generations, and check if they are independent. This study was carried out using articles collected from Scopus, within the field of psychology, which were divided into seven main subject categories:

(i) Applied psychology

(ii) Clinical psychology

(iii) Developmental and educational psychology

(iv) Experimental and cognitive psychology

(v) Neuropsychology and physiological psychology

(vi) Psychology (miscellaneous)

(vii) Social psychology

All articles were published in 2014 and their citation counts were collected in two time periods, giving two sets of data for the same sets of articles. The first set were gathered in November 2014, while the second were gathered in October 2016. Although articles published in November 2014 will have no time to attract any citation in the first set, whereas articles published in January 2014 will have up to 11 months to be cited, this is still a useful test to investigate the relationship between early and late citations.

The relationships were inspected using conditional probabilities. If we let G_1 represent the number of citations received in the first period, that is, the citation counts in November 2014, and G_2 be the number of citations received in the second period, that is, the difference in citation counts between the second and first set, then the relationship can be checked using:

(i) $P(G_2 > 0|G_1 = 0)$

(ii) $P(G_2 > 0|G_1 > 0)$

If these two conditional probabilities are equal, then it may be interpreted as implying that early citations do not influence later citations. On the other hand, if $P(G_2 > 0|G_1 > 0)$ is greater than $P(G_2 > 0|G_1 = 0)$, then this suggests that early citations do influence future citations. Otherwise, if $P(G_2 > 0|G_1 = 0) \approx 0$, then this suggests that a zero in the first generation usually leads to a zero in the second generation.

The results for this case study are given in Table 5.1. It is clear that in all cases, $P(G_2 > 0|G_1 > 0)$ are always greater than $P(G_2 > 0|G_1 = 0)$, implying that the generations are not completely independent. This supports the use of SVA and SVB models in citation analysis in the following sections.

Table 5.1: *Conditional probabilities based on two time periods for all investigated subjects.*

Subject	$P(G_2 > 0 \mid G_1 = 0)$	$P(G_2 > 0 \mid G_1 > 0)$
Applied psychology	0.733	0.876
Clinical psychology	0.712	0.862
Developmental and educational psychology	0.754	0.884
Experimental and cognitive psychology	0.811	0.889
Neuropsychology and physiological psychology	0.799	0.881
Psychology (miscellaneous)	0.657	0.881
Social psychology	0.704	0.850

5.3 Citation models with no covariates

Variants of compound models, especially the SVA, might come from a two generation process, where the first generation represents citations received shortly after a journal article has been published, and the second generation, perhaps overlapping with the first to some extent, represents the citations received as a result of scientists discovering an article because of its previous citations, either directly by following citations or indirectly because more cited articles are ranked more highly in some citation databases. SVB models may also be suitable due to the limitations of the citation database used to analyse the citations. For example, an article may be uncited in Scopus but cited in Google Scholar and its Google Scholar citations could attract new second generation citations.

5.3.1 Data and methods

Data from 20 different subject areas were selected from Scopus in order to assess the models for a wide range of different disciplines. This is important because citation patterns are known to vary considerably between disciplines. Thelwall and Wilson (2014a) analysed this data previously using the power law and discretised lognormal models. Each subject area is a single Scopus category and consists of all documents of type article that were published in 2004, giving ten years for the articles to attract citations. The total number of articles analysed for each subject area ranged from 528 to $5,000$ (see Table 5.5). When fitting the discretised lognormal model, one is added to all the counts as $\log(0)$ is undefined. In this section, we investigate:

1. Do PIG, ZIP, ZINB, Neyman type A, Polya Aeppli, and various SVA and SVB distributions fit citation count data better than discretised lognormal and negative binomial models?

2. If so, which model is preferable?

The models were fitted using R software (R Core Team, 2014), where the packages *gamlss* (Rigby et al., 2005), *MASS* (Venables and Ripley, 2002), *pscl* (Zeileis et al., 2008) were used to fit the PIG, negative binomial and zero-inflated models respectively. The codes given in Section 3.5 were used to fit the SVA and SVB models. We used an identity link and compared the models using AIC and BIC.

5.3.2 Model fitting results

The results (in terms of AIC and BIC) obtained when the models are fitted for each subject area are given in Tables 5.5, 5.6 and 5.7. Overall, the variants of compound models (SVA and SVB) produced a lower AIC and BIC compared to the Neyman type A, Polya Aeppli and PIG distributions. The SVB NB-NB model produced the lowest AIC for 13/20 subjects. The next most successful models are the SVA NB-NB and the discretised lognormal model. The SVB Poisson-NB and the SVB NB-Poisson each fitted best for only one subject (see Table 5.5). Adding the zero-inflated parameter to the Poisson distribution lowered its AIC for all subjects but this is not observed for the negative binomial model. The zero-inflated negative binomial (ZINB) model produced a higher AIC compared to the negative binomial model for all subjects except Ecology (see Table 5.6).

Similarly, the BIC also favoured the SVB NB-NB model as the BIC is lowest for 9/20 subjects. This is followed by the discretised lognormal and SVA NB-NB model, which has lowest BIC for 7/20 and 3/20 subjects respectively. Similarly to the AIC, the SVB-Poisson-NB is favoured by BIC for one subject, which is Rehab (see Table 5.7).

The estimated parameters for Tourism and Soil will be discussed because they are examples of subjects which return parameter estimates and errors for all the fitted distributions. From Table 5.2, when Tourism is fitted with the SVA Poisson-NB model, one generation follows the Poisson distribution with mean estimate $\lambda = 3.22$, whilst the other generation follows a negative binomial distribution with mean estimate $\mu = 18.77$ and size estimate $\alpha = 0.57$; thus the negative binomial generation has a variance of 640.19. However, when fitted with the SVA NB-Poisson model, the first generation follows a negative binomial distribution with mean estimate $\mu = 21.53$, size estimate $\alpha = 0.98$, and hence variance = 495.77, whilst the second generation follows a Poisson distribution

with mean estimate $\lambda = 0.01$. This latter estimate in practice indicates a negligible second generation. Moreover, the 95% confidence interval for λ_2 in Tourism contains zero, so the second Poisson generation here is insignificant. The estimated means (μ) in both negative binomial generations are relatively larger than the estimated means (λ) in the Poisson generations, suggesting that the majority of citation counts for Tourism derive from the negative binomial generation. This is consistent with the interpretation that the two generations occur simultaneously, instead of sequentially, as mentioned above. It is also interesting to note that the sum of the estimated means from the Poisson generations and negative binomial generations of these SVA models are approximately equal to the estimated mean when Tourism is fitted solely with the negative binomial model. When fitted with the SVA NB-NB model, the estimated mean for Tourism in the first NB generation (13.48) is larger than that of the second NB generation (8.25), suggesting that the majority of citation counts for Tourism derive from the first generation. Furthermore, the sum of the estimated means from the SVA NB-NB model for Tourism is also approximately equal to the estimated mean when Tourism is fitted with the negative binomial model only.

Similar results were obtained for Soil. When citation counts for Soil are fitted with the SVA Poisson-NB model and SVA NB-Poisson model, the mean estimates in the NB generations are much larger than those of the Poisson generations, suggesting that the majority of citation counts from Soil derive from the NB generation. Moreover, the sum of the estimated means for the SVA models are approximately equal to the estimated mean for the negative binomial model only (which is 16.93). It should be noted that the small estimates for λ_2 in the case of SVA NB-Poisson in Table 5.2 suggest that the second Poisson generation might not exist.

Table 5.2: *Estimated parameters for the negative binomial (NB), SVA Poisson-NB, SVA NB-Poisson and SVA NB-NB models.*

Sub.	NB		SVA Poisson-NB			SVA NB-Poisson			SVA NB-NB			
	μ	α	λ_1	μ_2	α_2	μ_1	α_1	λ_2	μ_1	α_1	μ_2	α_2
Tourism	21.5	1.0	3.2	18.8	0.6	21.5	1.0	0.0	13.5	1.3	8.3	0.1
Soil	16.9	0.7	2.3	16.1	0.6	16.9	0.7	0.1	13.8	0.8	3.5	0.0

Table 5.3 compares estimated parameters for the NB model against those of the SVB models. Similarly to the SVA distributions, Tourism and Soil depends largely on the generation that derives from the NB distribution, as the λ estimates are relatively smaller than the μ estimates. Furthermore, the sum of the two μ estimates for the SVB NB-NB models (21.533 and 16.931) are also similar to the

estimates from the NB distribution.

Table 5.3: *Estimated parameters for the NB, SVB Poisson-NB and SVB NB-NB models.*

Sub.	NB		SVB Poisson-NB			SVB NB-NB			
	μ	α	λ_1	μ_2	α_2	μ_1	α_1	μ_2	α_2
Tourism	21.5	1.0	1.4	20.1	0.8	14.8	0.4	6.8	1.2
Soil	16.9	0.7	0.1	16.8	0.7	4.9	0.1	12.0	0.8

Standard errors for SVA and SVB distributions

The associated standard errors for the estimated parameters for all subjects are given in Appendix C. Figures 5.1 and 5.2 show the mean and size estimates for the first and second generations of the SVB NB-NB model. Visual, Literature and Rehab were excluded as standard errors could not be obtained as a result of singular hessian matrices.

Although the SVB NB-NB model has the lowest AIC/BIC for most subjects, the model returned very large standard errors, resulting in large confidence intervals, as shown in Figures 5.1 and 5.2, indicating that this SVB NB-NB model is impractical, as the parameter estimates will be less precise. This result could possibly be due to the nature of citations differing from that of the larvae studied by Neyman. With larvae and their offspring it is clear which generation of population a larvae originates from but this is not the case with citations. Usually it will be far from clear cut which generation a given citation might belong to, which in turn leads to difficulties estimating the mean number of citations for that generation, and hence the large associated standard errors.

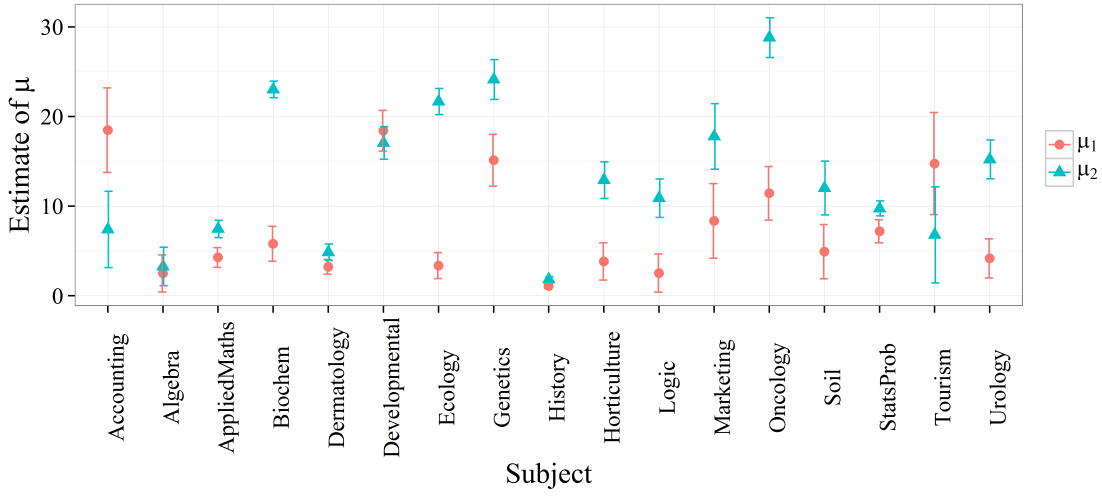

Figure 5.1: *Mean (μ) estimates of the SVB NB-NB model for first and second generations with* 95% *confidence intervals*

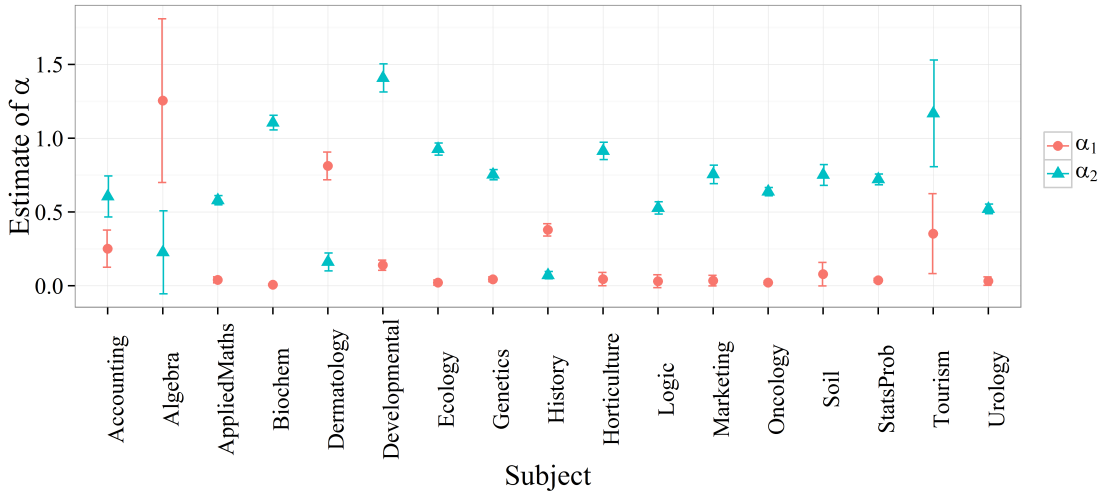

Figure 5.2: *Size (α) estimates of the SVB NB-NB model for first and second generations with 95% confidence intervals*

Further simulation studies were carried out using the SVA and SVB models to check if similar error patterns are obtained. In each case, 1000 data points are simulated using known fixed parameters, selected at random, and refitted to the models. These cases are each repeated 2000 times. For each repetition, the estimated parameters are recorded. The associated standard errors are obtained using the standard deviation of the estimated parameters. A summary of the results obtained is in Table 5.4.

Table 5.4: *Results obtained when simulated SVA or SVB data are refitted to the simulation model. The presented estimated parameters are means from* 2000 *repetitions.*

Simulation model	Coefficients		Standard errors
SVA Poisson-NB$(2, 3, 1)$	λ	2.003	0.079
	μ	2.995	0.140
	α	1.009	0.104
SVA NB-Poisson$(3, 1, 2)$	μ	2.99	0.139
	α	1.003	0.067
	λ	2.004	0.123
SVA NB-NB$(10, 3, 8, 1)$	μ_1	10.234	1.879
	α_1	6.636	65.204
	μ_2	7.755	1.914
	α_2	1.200	4.289
SVB Poisson-NB$(2, 3, 1)$	λ	2.025	0.014
	μ	3.039	0.039
	α	1.048	0.025
SVB NB-NB$(5, 3, 4, 2)$	μ_1	5.086	1.785
	α_1	97.414	925.658
	μ_2	3.913	1.788
	α_2	4.092	47.753

Table 5.4 shows that the α_1 estimate of both the SVA NB-NB and SVB NB-NB vary largely, with large associated error. Hence it can be concluded that both the SVA NB-NB and SVB NB-NB are impractical when modelling data with no covariates.

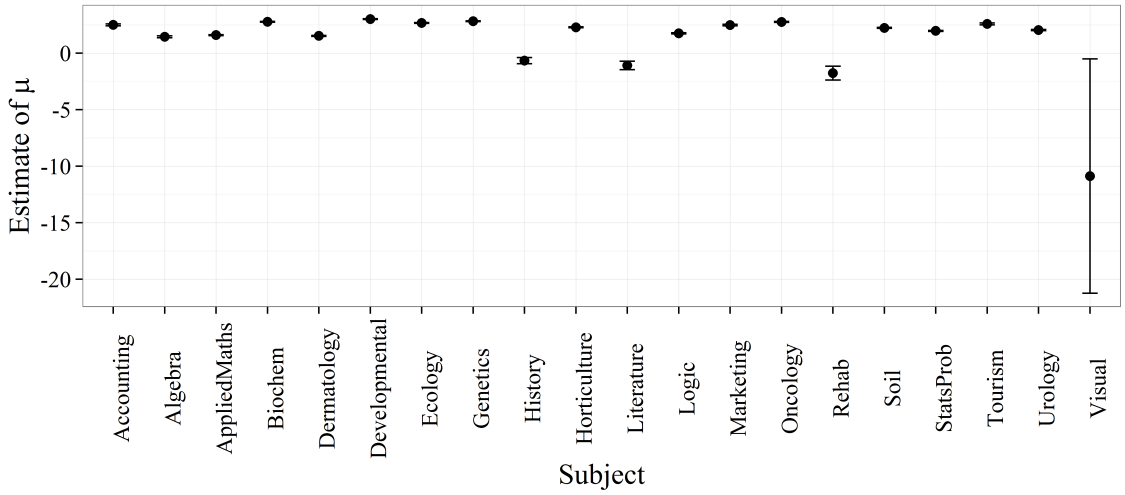

Figure 5.3: *Log of the mean (μ) estimates for the discretised lognormal distribution with* 95% *confidence intervals*

On the other hand, the 95% confidence intervals for all subjects except Visual

for the discretised lognormal distribution (Figure 5.3) are much narrower compared to those for the SVB NB-NB model. This indicates that the discretised lognormal distribution is more suitable in practice when there are no covariates.

Randomised quantile residual plots

The models are further examined using randomised quantile residual plots. In this section, the plots for *Tourism* will be discussed in detail, the plots for the other subjects are given in Appendix E. Dunn and Smyth (1996) suggest plotting four realisations of the quantile residuals to account for the randomisation, in case some inconsistent patterns are present, but only one realisation is presented here as the patterns are preserved.

The randomised quantile residual plot obtained when the negative binomial model is fitted to Tourism is given in Figure 5.4. The plotted points are roughly on the line, indicating that the negative binomial model is a good fit. The last outstanding point indicate that the largest observed count of 257, is bigger than that expected under the negative binomial model. However, this is negligible as it refers to the 99.9^{th} percentile, since $P(Z < 3) = 0.9987$.

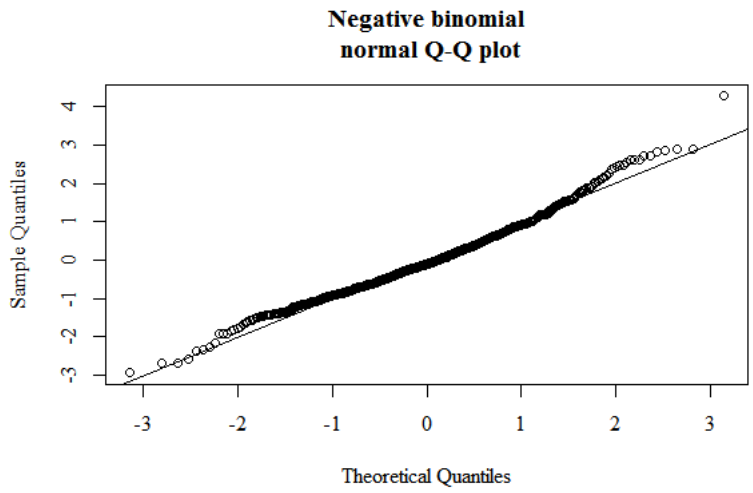

Figure 5.4: *Randomised quantile residual plot for Tourism when fitted with the negative binomial model.*

The randomised quantile residual plots for discretised lognormal, SVA and SVB models are given in Figure 5.5. The plot for the discretised lognormal model shows that for values that are approximately after the 97^{th} percentile, (since $P(Z < 2) = 0.977$), which refers to citation counts in the range of $108 - 257$, the observed counts are less than the expected under the fitted discretised lognormal model. However, since this only accounts for about the top 2% of data, the model

is still adequate. Overall, the plots in Figure 5.5 indicate that the fitted SVA and SVB models are also adequate.

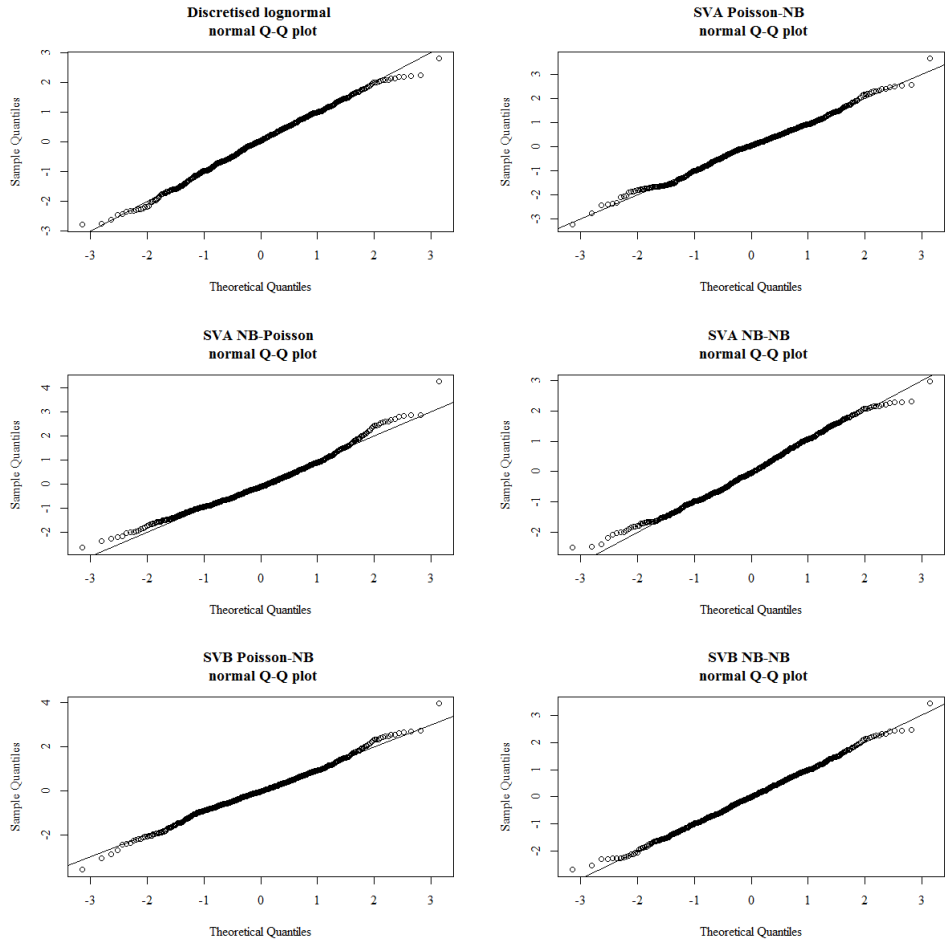

Figure 5.5: *Randomised quantile residual plots for Tourism when fitted with discretised lognormal, SVA and SVB models.*

5.3.3 Discussion and summary

The large standard errors in the SVB models could be due to the increased flexibility of the model, because SVB models can have a zero in the first generation followed on by a non-zero in the second generation, increasing the possible variation in the distribution. This scenario is possible for citations, albeit perhaps rare. For example, 'Sleeping Beauties' attract many citations after a long period without any (Van Raan, 2004). The variant models, especially SVA, replicate to some extent the rich get richer or Matthew effect introduced by Merton (1968), because more citations in the first generation tend to lead to more citations in the second generation. However, whilst the Matthew effect is an ongoing process, the SVA increase occurs only in a single leap, in the transition from the first to second generation.

Although the SVB NB-NB model produced the lowest AIC and BIC in many cases, it is also associated with large standard errors of parameter estimates, and it could be argued that the large standard errors render this model impractical. On the other hand, the discretised lognormal model tends to have good AIC and BIC (being the next most successful model), and the parameter estimates nearly always have smaller errors. However, due to the nature of the different models, this direct comparison of standard errors can be misleading. Given that the use of SVB models in citation analysis is uncommon, and the property of having large associated standard errors in the parameter estimates outweighs the low AIC/BIC, this may be of limited interest to the community. On the other hand, SVA models tend to have smaller standard errors than SVB, and are more practical, and thus SVA models are recommended for future citation analysis as a possible alternative to the discretised lognormal. On a theoretical level, the good fits found for some of the proposed variants of compound models give evidence that there may be (at least) two important and separate processes that govern the citing practices of authors. For one of these processes, existing citations are irrelevant for new citations, and for the other, they are relevant. In both circumstances, a larger mean in the first generation will on average, lead to a larger mean in the second generation. This property demonstrated by the variants of compound models, which is a partial rich-get-richer effect, indicates that an article might have low citation counts because it missed out on the first generation. Therefore, the two processes highlight the importance of early publicity for research. In other words, if authors publicise their research during the early stage, then this initial interest may attract more citations in the second generation.

Table 5.5: *AIC for all subjects when fitted with discretised lognormal, negative binomial and variants of compound models (the lowest AIC produced by models for each subject is in bold).*

Subjects	Discretised lognormal	Negative binomial	SVA Poisson-NB	SVA NB-Poisson	SVA NB-NB	SVB Poisson-NB	SVB NB-NB	Number of articles
Visual	7,902	7,928	7,916	7,930	7,865	7,920	**7,865**	4,096
Tourism	4,956	4,980	4,980	4,982	4,969	4,964	**4,955**	608
Soil	33,470	33,344	33,458	33,345	33,287	33,344	**33,282**	4,347
Marketing	**12,917**	13,073	13,025	13,073	12,941	13,015	12,932	1,550
Literature	11,624	11,635	**11,618**	11,637	11,622	104,485	25,449	5,000
Horticulture	23,058	23,093	23,165	23,095	23,001	23,067	**22,992**	3,009
History	19,797	19,994	19,849	19,996	19,824	19,880	**19,795**	5,000
Genetics	45,622	46,014	45,997	46,002	45,474	45,982	**45,471**	5,000
Ecology	42,787	42,343	42,441	42,335	42,253	42,366	**42,240**	5,000
Developmental	40,985	41,604	41,340	41,558	40,979	41,385	**40,956**	4,541
Biochem	42,901	43,690	43,540	43,638	**42,675**	43,659	42,680	5,000
Accounting	9,927	9,933	9,924	9,931	9,914	9,929	**9,896**	1,178
AppliedMaths	33,504	33,739	33,704	33,741	33,460	33,685	**33,441**	5,000
Urology	38,932	38,621	38,793	38,623	**38,560**	38,623	38,563	5,000
StatsProb	**36,696**	37,416	37,177	37,418	36,742	37,186	36,706	5,000
Rehab	28,086	27,531	27,622	27,533	27,628	**27,483**	28,322	5,000
Oncology	42,577	42,620	42,679	42,607	**42,196**	42,660	42,225	4,646
Logic	32,258	32,044	32,164	32,046	32,012	32,045	**32,010**	4,547
Dermatology	19,608	19,774	19,671	19,776	19,675	19,692	**19,606**	3,184
Algebra	**2,968**	2,991	2,973	2,993	2,977	2,978	2,972	528

Table 5.6: *AIC for all subjects when fitted with negative binomial, Poisson, Neyman type A, Polya Aeppli, Poisson Inverse Gaussian (PIG), ZIP and ZINB.*

Subjects	Discretised lognormal	Negative binomial	Poisson	Neyman type A	Polya Aeppli	PIG	ZIP	ZINB
Visual	***7,902**	7,928	13,077	8,534	8,078	7,923	9,720	7,931
Tourism	***4,956**	4,980	16,268	5,812	5,123	4,967	15,757	4,983
Soil	33,470	***33,344**	107,239	39,120	34,131	33,760	96,159	33,347
Marketing	**12,917**	13,073	76,697	NA	13,817	13,007	72,490	13,075
Literature	***11,624**	11,635	15,615	12,061	11,753	11,637	12,825	11,638
Horticulture	***23,058**	23,093	72,451	32,007	23,746	23,193	67,339	23,095
History	***19,797**	19,994	47,108	23,669	20,877	19,847	35,450	19,997
Genetics	***45,622**	46,014	372,128	NA	79,913	46,199	350,902	46,016
Ecology	42,787	42,343	152,989	62,755	42,833	43,336	137,563	***42,292**
Developmental	***40,985**	41,604	220,499	NA	43,242	41,097	216,794	41,607
Biochem	***42,901**	43,690	286,131	NA	45,616	43,361	276,641	43,692
Accounting	***9,927**	9,933	48,202	13,318	10,310	10,018	44,042	9,935
AppliedMaths	***33,504**	33,739	127,994	69,346	35,349	33,805	109,743	33,741
Urology	38,932	***38,621**	171,536	54,326	39,831	39,584	144,252	38,624
StatsProb	**36,696**	37,416	195,493	NA	39,992	36,950	179,399	37,419
Rehab	28,086	***27,531**	126,891	39,065	28,697	28,443	83,243	27,533
Oncology	***42,577**	42,620	363,127	NA	44,562	43,416	330,526	42,622
Logic	32,258	***32,044**	110,369	38,902	32,931	32,680	91,249	32,047
Dermatology	***19,608**	19,774	49,322	23,020	20,482	19,661	43,356	19,777
Algebra	**2,968**	2,991	5,261	3,237	3,057	2,970	4,783	2,994

*Lowest AIC produced by models in Table 5.6 for each subject area, but an even lower AIC can be found for respective subject areas using models as in Table 5.5.
In no cases is the Neyman type A model the superior model. "NA" has been placed for cases where the distribution failed, indicating that the Neyman type A is completely unsuitable.

Table 5.7: *BIC for all subjects when fitted with models (the lowest BIC produced by models for each subject is in bold).*

Subjects	Discretised lognormal	Negative binomial	SVA Pois-NB	SVA NB-Pois	SVA NB-NB	SVB Pois-NB	SVB NB-NB	Neyman type A	Polya Aeppli	PIG	ZIP	ZINB
Visual	7,915	7,940	7,935	7,949	7,891	7,939	**7,890**	8,546	8,091	7,936	9,733	7,951
Tourism	**4,965**	4,989	4,993	4,996	4,987	4,977	4,972	5,820	5,132	4,976	15,766	4,998
Soil	33,483	33,357	33,477	33,364	33,313	33,363	**33,308**	39,132	34,143	33,773	96,172	33,366
Marketing	**12,928**	13,083	13,041	13,089	12,962	13,031	12,953	NA	13,828	13,018	72,501	13,094
Literature	**11,637**	11,649	11,637	11,657	11,648	104,504	25,475	12,074	11,766	11,650	12,838	11,658
Horticulture	23,070	23,105	23,183	23,113	23,025	23,085	**23,016**	32,019	23,758	23,205	67,351	23,114
History	**19,810**	20,007	19,868	20,016	19,850	19,900	19,821	23,682	20,890	19,860	35,463	20,016
Genetics	45,635	46,027	46,016	46,022	45,500	46,001	**45,497**	NA	79,926	46,212	350,915	46,081
Ecology	42,800	42,356	42,461	42,354	42,279	42,386	**42,266**	62,768	42,846	43,349	137,576	42,311
Developmental	40,998	41,617	41,359	41,577	41,005	41,404	**40,982**	NA	43,255	41,109	216,807	41,625
Biochem	42,914	43,703	43,559	43,657	**42,702**	43,679	42,706	NA	45,629	43,374	276,654	43,712
Accounting	9,937	9,943	9,940	9,946	9,934	9,944	**9,917**	13,328	10,320	10,028	44,053	9,954
AppliedMaths	33,517	33,752	33,723	33,761	33,486	33,705	**33,467**	69,359	35,362	33,819	109,756	33,761
Urology	38,945	38,634	38,812	38,643	**38,586**	38,642	38,589	54,339	39,844	39,597	144,265	38,643
StatsProb	**36,709**	37,429	37,196	37,437	36,768	37,205	36,732	NA	40,005	36,963	179,412	37,437
Rehab	28,099	27,544	27,641	27,553	27,654	**27,503**	28,348	39,078	28,711	28,456	83,256	27,567
Oncology	42,590	42,633	42,698	42,626	**42,222**	42,680	42,251	NA	44,575	43,429	330,538	42,703
Logic	32,271	32,057	32,183	32,066	32,037	32,064	**32,036**	38,915	32,943	32,693	91,262	32,066
Dermatology	**19,620**	19,787	19,690	19,795	19,699	19,710	19,631	23,032	20,494	19,674	43,368	19,795
Algebra	**2,976**	3,000	2,986	3,006	2,994	2,991	2,990	3,245	3,066	2,978	4,792	3,006

"NA" has been placed for cases where the distribution failed, indicating that the model is completely unsuitable.

5.4 Citation analysis with covariates

Motivations and factors that influence citing practises are of interest to help understand the influences behind citations and hence how to interpret citation counts (Case and Higgins, 2000; Didegah and Thelwall, 2013a). Vieira and Gomes (2010) showed that the number of co-authors and affiliations has an effect on the citation rate of a paper. Here, we incorporated two other covariates, number of authors and number of affiliated countries (i.e., the number of countries listed in the addresses of the authors, reflecting the degree of internationality of any collaboration), into our models. Both were extracted from the Scopus records of the articles. The correlation between the two covariates is low, hence there is no problem with collinearity.

5.4.1 Data and methods

The citation data analysed in this section was extracted from Scopus in 2015, and consists of counts of citations to journal articles published in 2009 from 24 different subject areas (see Table 5.8). The time frame, also known as a citation window, gives articles six years to attract citations, allowing us to obtain a more stable pattern. It is also vital to incorporate different areas as citation behaviour varies across disciplines (Leydesdorff, 2013). For example, articles in medical journals tend to receive more citations than the articles in social science and humanities journals (Rauhvargers, 2011). Citation counts are therefore only comparable within a field, but not between fields.

Table 5.8: *The subjects investigated and their name abbreviations*

Original subject names	Abbreviated subject names	No. of articles
Applied Mathematics	Applied maths	6,411
Aquatic Science	Aquatic	9,123
Archeology (arts and humanities)	Archeology	1,212
Biochemistry (medical)	Biochemistry	3,279
Biomedical Engineering	Biomedical	4,796
Biophysics	Biophysics	8,601
Care Planning	Care planning	266
Cellular and Molecular Neuroscience	Neuroscience	7,476
Chemical Health and Safety	Chemical health	243
Computer Graphics and Computer Aided Design	Computer	3,537
Condensed Matter Physics	Physics	5,810
Developmental and Educational Psychology	Developmental	7,338
Earth Surface Processes	Earth	6,188
Education	Education	6,937
Electronic Optical and Magnetic Materials	Electronic	4,826
Environmental Chemistry	Environmental	8,839
Inorganic Chemistry	Inorganic	9,330
Management Information Systems	Management	1,830
Microbiology	Microbiology	8,701
Nuclear Energy and Engineering	Nuclear	5,300
Oral Surgery	Oral surgery	389
Pharmacology	Pharmacology	3,769
Small Animals	Small animals	1,063
Statistics Probability and Uncertainty	Statistics	5,011

In this section, we compare the fits of negative binomial, Neyman type A, Polya Aeppli, SVA and SVB models. We model all parameters using a log-link and the quadratic model:

$$\textit{Citation count} \sim \textit{Number of authors} + \textit{Number of affiliations} \tag{5.1}$$

The NB size parameter is assumed to be constant. All models were fitted using R (R Core Team, 2014) and parameters were estimated using maximum likelihood estimations via the *optim* function using code presented in Section 3.6. We fitted the standard negative binomial models using the MASS (Venables and Ripley, 2002) and gamlss (Rigby et al., 2005) packages in R. Note that when fitting the standard negative binomial model in gamlss, the variance function is $\mu + \mu^2\alpha$.

We made slight modifications to the code used by Oliveira et al. (2016) to fit the Neyman type A and Polya-Aeppli models. In addition, models were compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

5.4.2 Results

The results in terms of log-likelihood, AIC and BIC are presented in Tables 5.9, 5.10 and 5.11. We found that both AIC and BIC produced similar results by selecting the same superior models for most subjects, except for 6 subjects, which are Biophysics, Cellular and Molecular Neuroscience, Computer Graphics and Computer Aided design, Inorganic Chemistry, Management Information Systems, and Microbiology. Although the SVB models are more flexible, it is clear that the SVA models are better in terms of log-likelihood, AIC and BIC. The SVA NB-NB model was favoured by AIC and BIC in 12/24 and 8/24 subjects respectively, while BIC also favoured the negative binomial model in 8/24 cases.

Whilst opinions differ about the minimum difference between AICs to be significant, in 14/24 models, the AIC of the 'superior' model was at least 23 less than the next best model, showing that the winner is clear cut.

The estimated coefficients for the fitted models are given in Appendix D. Focusing on the estimated coefficients for the SVA NB-NB model (see Table 5.12), in most subjects, the number of affiliated countries has a greater impact on the estimated mean than does the number of authors. The estimated coefficients for the parameters of the SVA NB-NB model confirms the findings of Nomaler et al. (2013) that the number of affiliated countries has a bigger impact on the number of citations received compared to the number of authors.

Table 5.9: *Log-likelihood for the Neyman type A, Polya Aeppli, negative binomial, SVA and SVB models*

Subjects	Neyman type A	Polya Aeppli	NB	SVA Pois-NB	SVA NB-Pois	SVA NB-NB	SVB Pois-NB	SVB NB-NB
Applied maths	−29,897	−20,805	−19,847	−19,869	−19,847	**−19,748**	−19,846	−19,832
Aquatic	−33,194	−31,284	−30,948	−31,077	−30,954	**−30,885**	−30,944	−30,948
Archeology	−2,508	−2,380	**−2,285**	−2,292	−2,285	−2,288	−2,285	−2,317
Biochemistry	NA	−16,859	**−11,714**	−11,745	−11,714	−11,849	−11,722	−11,795
Biomedical	−31,188	−17,351	−16,917	−17,003	−16,923	**−16,864**	−16,917	−16,952
Biophysics	−38,187	−33,286	−32,421	−32,788	−32,425	−33,324	−32,422	**−32,413**
Care planning	−769	−704	−676	−682	**−676**	−678	NA	−690
Neuroscience	−37,871	−30,172	−29,718	−29,830	−29,666	**−29,664**	NA	−29,725
Chemical health	−812	−761	−736	**−721**	−733	−747	−737	−736
Computer	−13,511	−11,980	−11,033	−11,030	NA	**−11,017**	−11,033	−11,033
Physics	NA	−33,019	−20,152	**−20,013**	−20,154	−20,189	NA	−20,070
Developmental	−29,632	−26,672	−26,273	−26,572	−26,257	**−26,180**	−26,264	−26,263
Earth	−23,167	−21,188	−21,190	−21,318	**−21,187**	−21,274	−21,190	−21,190
Education	−23,964	−20,836	−20,108	−21,096	−20,108	**−20,078**	−20,122	−20,108
Electronic	−22,049	−18,581	−17,853	**−17,822**	−17,859	−17,850	−17,842	−17,847
Environmental	−44,651	−35,453	−34,778	−34,863	−34,777	**−34,717**	−34,778	−34,779
Inorganic	−38,202	−34,207	−33,464	**−33,460**	−33,464	−33,496	−33,469	−33,464
Management	−7,448	−6,235	−5,972	−5,977	−5,971	**−5,964**	−5,972	−5,972
Microbiology	−35,881	−33,124	−32,814	−33,294	−32,767	**−32,764**	−32,821	−32,815
Nuclear	−20,298	−16,539	−15,941	−15,955	−15,941	**−15,890**	NA	−15,941
Oral Surgery	−1,341	−1,294	−1,280	−1,280	**−1,280**	−1,287	−1,280	−1,287
Pharmacology	−15,672	−12,638	−12,464	**−12,421**	−12,464	−12,644	−12,463	−12,505
Small Animals	−3,022	**−2,834**	−2,949	−3,049	−2,949	−2,932	−2,964	−2,983
Statistics	−18,455	−16,017	−15,285	−15,324	−15,292	**−15,078**	−15,282	−15,285

Note: The model which give the highest log-likelihood value is in bold. "NA" indicates that the model is inappropriate.

Table 5.10: *AIC for the Neyman type A, Polya Aeppli, negative binomial, SVA and SVB models*

Subjects	Neyman type A	Polya Aeppli	NB	SVA Pois-NB	SVA NB-Pois	SVA NB-NB	SVB Pois-NB	SVB NB-NB
Applied maths	59,803	41,618	39,701	39,752	39,707	**39,513**	39,706	39,680
Aquatic	66,395	62,576	61,904	62,167	61,923	**61,787**	61,902	61,912
Archeology	5,024	4,768	**4,579**	4,598	4,585	4,592	4,585	4,650
Biochemistry	NA	33,727	**23,435**	23,503	23,443	23,714	23,458	23,606
Biomedical	62,385	34,711	33,842	34,019	33,861	**33,743**	33,849	33,921
Biophysics	76,382	66,581	64,849	65,590	64,865	66,665	64,858	**64,842**
Care planning	1,545	1,416	**1,361**	1,378	1,366	1,372	NA	1,395
Neuroscience	75,750	60,353	59,443	59,673	59,346	**59,344**	NA	59,467
Chemical health	1,631	1,529	1,479	**1,456**	1,480	1,511	1,489	1,487
Computer	27,030	23,967	22,075	22,073	NA	**22,050**	22,081	22,083
Physics	NA	66,046	40,311	**40,040**	40,322	40,393	NA	40,155
Developmental	59,272	53,352	52,553	53,159	52,529	**52,375**	52,542	52,542
Earth	46,341	**42,385**	42,388	42,649	42,388	42,564	42,394	42,396
Education	47,936	41,681	40,225	42,206	40,231	**40,173**	40,258	40,233
Electronic	44,106	37,169	35,714	**35,658**	35,733	35,716	35,699	35,709
Environmental	89,310	70,914	69,564	69,740	69,567	**69,450**	69,570	69,573
Inorganic	76,411	68,421	66,937	66,934	**66,943**	67,007	66,953	66,945
Management	14,903	12,479	11,951	11,969	11,955	**11,944**	11,957	11,959
Microbiology	71,770	66,257	65,636	66,603	65,548	**65,544**	65,657	65,645
Nuclear	40,603	33,086	31,890	31,923	31,896	**31,796**	NA	31,899
Oral surgery	2,689	2,596	**2,569**	2,574	2,573	2,590	2,575	2,589
Pharmacology	31,352	25,283	24,936	**24,855**	24,943	25,303	24,941	25,027
Small animals	6,053	**5,676**	5,906	6,112	5,912	5,880	5,941	5,981
Statistics	36,917	32,042	30,577	30,662	30,597	**30,173**	30,578	30,585

Note: The model which give the lowest AIC is in bold. "NA" indicates that the model is inappropriate.

Table 5.11: *BIC for the Neyman type A, Polya Aeppli, negative binomial, SVA and SVB models*

Subjects	Neyman type A	Polya Aeppli	NB	SVA Pois-NB	SVA NB-Pois	SVA NB-NB	SVB Pois-NB	SVB NB-NB
Applied maths	59,830	41,645	39,728	39,799	39,755	**39,567**	39,754	39,735
Aquatic	66,424	62,605	61,932	62,217	61,972	**61,844**	61,952	61,969
Archeology	5,044	4,788	**4,599**	4,634	4,620	4,633	4,620	4,691
Biochemistry	NA	33,751	**23,460**	23,546	23,486	23,763	23,501	23,655
Biomedical	62,410	34,737	33,868	34,065	33,906	**33,795**	33,894	33,973
Biophysics	76,410	66,609	**64,878**	65,640	64,914	66,721	64,908	64,899
Care planning	1,560	1,430	**1,375**	1,404	1,391	1,401	NA	1,424
Neuroscience	75,778	60,381	59,471	59,722	**59,395**	59,399	NA	59,522
Chemical health	1,645	1,543	1,493	**1,481**	1,505	1,539	1,513	1,515
Computer	27,055	23,992	**22,099**	22,117	NA	22,100	22,124	22,132
Physics	NA	66,072	40,338	**40,087**	40,369	40,446	NA	40,209
Developmental	59,299	53,379	52,581	53,207	52,577	**52,430**	52,590	52,598
Earth	46,368	**42,412**	42,415	42,696	42,436	42,617	42442	42,449
Education	47,963	41,708	40,252	42,254	40,279	**40,228**	40,306	40,287
Electronic	44,132	37,195	35,740	**35,703**	35,778	35,767	35,744	35,761
Environmental	89,339	70,943	69,593	69,789	69,617	**69,506**	69,620	69,630
Inorganic	76,440	68,450	**66,965**	66,984	66,993	67,065	67,003	67,002
Management	14,925	12,501	**11,973**	12,007	11,994	11,988	11,996	12,003
Microbiology	71,798	66,285	65,665	66,652	**65,597**	65,600	65,706	65,702
Nuclear	40,630	33,113	31,916	31,969	31,942	**31,848**	NA	31,952
Oral Surgery	2,705	2,612	**2,584**	2,602	2,601	2,622	2,603	2,621
Pharmacology	31,376	25,308	24,961	**24,899**	24,986	25,353	24,984	25,076
Small Animals	6,073	**5,696**	5,926	6,146	5,946	5,920	5,976	6,021
Statistics	36,943	32,068	30,603	30,707	30,643	**30,225**	30,624	30,637

Note: The model which give the lowest BIC is in bold. "NA" indicates that the model is inappropriate.

Table 5.12: *Estimated coefficients when citation data are fitted with the SVA NB-NB model.*

Subject	Estimated coefficients							
	μ_{inter1}	μ_{auth1}	μ_{coun1}	α_1	μ_{inter2}	μ_{auth2}	μ_{coun2}	α_2
Applied maths	1.20	0.18	0.32	−0.47	0.11	−0.53	3.73	−6.27
Aquatic	0.97	0.11	0.37	0.42	2.15	−0.02	−0.44	−1.56
Archeology	0.21	0.19	0.19	−0.99	5.76	−1.22	−1.76	−6.40
Biochemistry	0.67	0.05	−0.03	2.04	0.83	0.19	0.32	−0.76
Biomedical	−0.75	0.08	0.87	1.55	2.50	0.09	−0.13	−0.74
Biophysics	2.22	0.05	−0.90	1.21	1.12	0.15	0.61	−0.58
Care planning	−0.33	0.17	0.02	2.75	0.04	0.36	0.25	−0.72
Neuroscience	2.37	0.01	−0.04	0.74	2.47	−0.02	0.11	−1.62
Chemical health	2.44	−0.10	0.04	0.06	0.98	0.33	−4.33	4.51
Computer	1.06	0.00	0.90	−0.67	2.16	−1.92	5.04	−5.68
Physics	0.56	−0.03	0.27	3.49	1.38	−0.02	0.94	−0.94
Developmental	1.52	0.12	0.31	−0.08	2.30	0.14	−1.36	−2.54
Earth	1.57	0.20	−0.02	−0.21	2.03	−1.97	1.95	−2.56
Education	0.95	0.27	0.31	−0.64	9.38	−4.87	5.48	−6.88
Electronic	0.67	0.05	−0.02	2.50	2.27	0.03	0.20	−0.70
Environmental	2.48	0.06	0.12	−0.02	6.39	−0.15	−1.48	−7.07
Inorganic	0.41	0.20	0.30	0.40	4.21	−0.18	−1.15	−1.37
Management	0.03	0.16	0.26	1.81	0.67	0.01	1.54	−1.12
Microbiology	2.19	0.05	0.15	−0.12	−0.15	−0.14	0.66	−0.31
Nuclear	1.68	0.03	0.15	−0.68	2.14	−0.34	2.46	−6.46
Oral Surgery	1.43	0.04	0.47	0.45	5.11	−1.30	1.08	−3.78
Pharmacology	0.80	0.06	−0.37	3.83	2.37	0.00	0.16	−0.81
Small Animals	0.08	0.40	−0.18	−0.19	2.45	−0.80	1.23	−1.09
Statistics	1.18	0.16	0.04	−0.05	−1.15	0.24	1.71	−3.12

Randomised quantile residual plots

The models are further examined using randomised quantile residual plots. In this section, the plots for *Applied Mathematics* and *Aquatic Science* will be discussed in detail. The plots for the other subjects are given in Appendix F.

Applied Mathematics

The randomised quantile residual plot obtained when negative binomial model is fitted to Applied Mathematics is in Figure 5.6. The plot shows that there are large positive residuals, for values that are approximately after the 97^{th} percentile (since $P(Z < 2) = 0.977$), which refer to citation counts above 50, the observed quantile residuals are larger than those expected under the standard normal distribution. This indicates that the large citation counts (in the range of $50 - 743$) for Applied Mathematics are larger than expected under the fitted negative binomial model. However, as this only applies for relatively small number of data thus this model is still adequate.

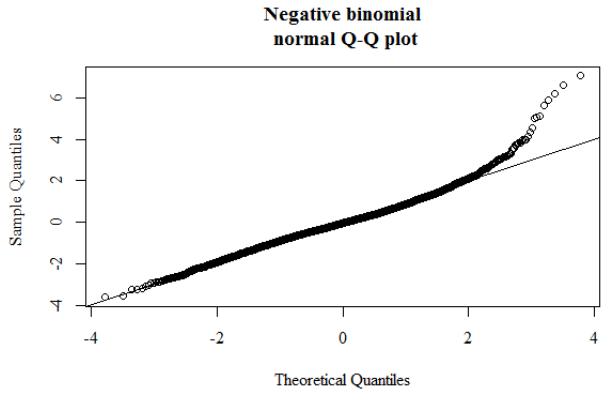

Figure 5.6: *Randomised quantile residual plot when Applied Mathematics are fitted with the negative binomial model.*

The randomised quantile residual plots for the SVA and SVB models are given in Figure 5.7. Similar to the negative binomial model, all models, apart from SVA NB-NB, produce large positive residuals. Comparing all the plots in Figure 5.7, it is clear that the SVA NB-NB model is the most suitable.

It is noted that the randomised quantile residual plots compare the cumulative distribution of the observed data against those expected from the null model. This neglects the individual counts as an excess count may be counterbalanced if only few are present thereafter. For example, under the null model, whilst an excess number of zeros are present in the data, the observed number of ones are less than expected, if the number of observations of the former counterbalances the latter, then the cumulative distribution will still be consistent with the null model. It is difficult to detect this using the randomised quantile residual plots. Hence an

alternative method which take individual counts into consideration is discussed in Chapter 6.

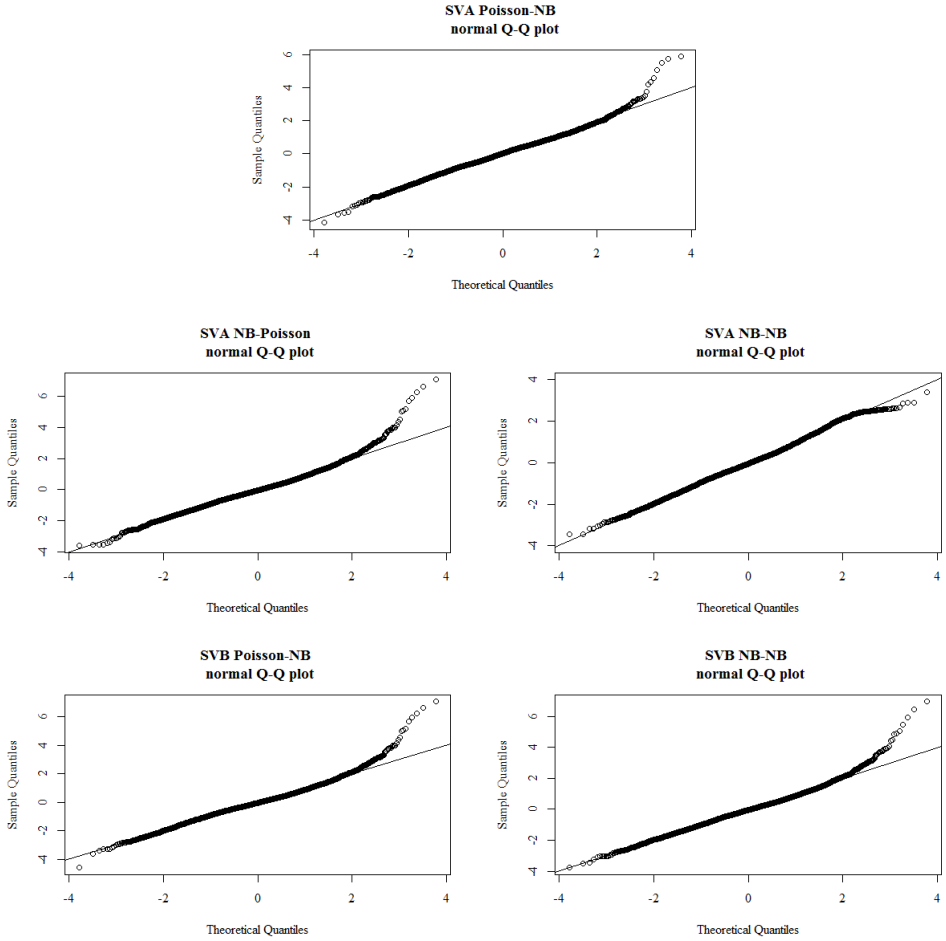

Figure 5.7: *Randomised quantile residual plots for Applied Mathematics when fitted with the SVA and SVB models.*

Aquatic Science

When the negative binomial model is fitted to Aquatic Science, the associated randomised quantile residual plot in Figure 5.8 is obtained. Here, the sample and theoretical quantiles are consistent up to about the 97^{th} percentile, which are citation counts less that 42. Beyond this and up to 374, large positive residuals are present, indicating that the observed counts (between $42 - 374$) are greater than the expected under the negative binomial model.

The randomised quantile residual plots for Aquatic Science when fitted with the SVA and SVB models are given in Figure 5.9. Similar to the negative binomial model, large positive residuals are observed after approximately the 97^{th} percentile for all other models, except for SVA NB-NB, which is the most suitable.

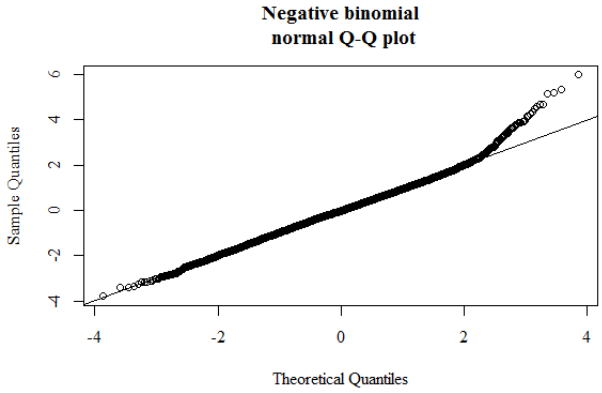

Figure 5.8: *Randomised quantile residual plot when Aquatic Science are fitted with the negative binomial model.*

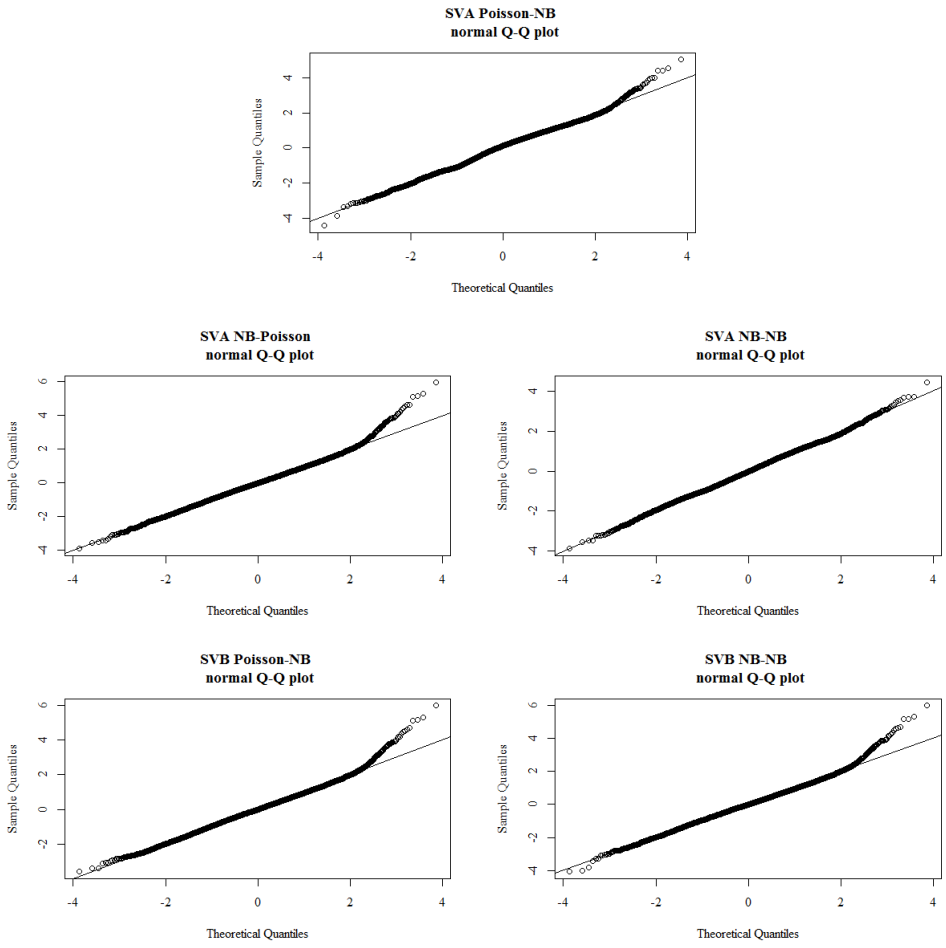

Figure 5.9: *Randomised quantile residual plot when Aquatic Science are fitted with the SVA and SVB models.*

5.4.3 Discussion and summary

We compared our proposed variants of compound models to other standard models such as negative binomial, and obtained mixed results. In rare cases, "NA"

is obtained, indicating that the model is inappropriate for that subject. Further studies on the reasons for this occurrence may be beneficial, however this is not performed in this thesis as the models provide a good fit in general.

The suitability of some of the proposed variants of compound models in our results gave evidence that there are (at least) two important processes governing the citing practises of authors. Our results also show that the number of affiliated countries has a greater impact than that of number of authors, which is consistent with previous findings. In addition, the estimated coefficients obtained in the second NB generation are larger compared to those in the first NB generation in SVA NB-NB models, which is consistent with the 'rich get richer' effect, as the initial interest/citations leads to more second generation citations. Nevertheless, due to the varying citation behaviour across different fields, it is important to note that our results are subject specific, and thus cannot be used to directly compare subjects across different fields.

5.5 Biodosimetry analysis

The field of biodosimetry is important especially in long term health risk studies, as it involves measurements of biological markers, such as frequency of chromosome aberration, following radiation exposure (Simon et al., 2010). Research generally focus on the identification of appropriate distributions for the cytogenetic dose-response curve (Oliveira et al., 2016), that is, to estimate the initial response distribution, given some radiation doses. Often only the frequency of abberrated chromosomes is observed, without the knowledge of exposed radiation doses. Hence, having an adequately correct response distribution will aid in assessing patients quicker in the case of large scale radiation accidents (Romm et al., 2013). It is common for the response variable to be in the form of counts, and hence count models such as compound models, and particularly compound Poisson models are commonly used in this field. For example, Virsik and Harder (1981) and Puig and Barquinero (2011) classed the first generation as the particles traversing a cell nucleus, and the second generation as the number of dicentrics, which are abnormal chromosomes with two centromeres, produced by each of these particles.

The analysis in this section includes the application of SVA and SVB models to biodosimetry data. Although there are no rationale interpretation when applying these models, as a dicentric chromosome is unlikely to influence its neighbouring chromosome, it may be useful to compare the use of compound models to their variants. This analysis may still be used as an illustrative example, where models are used without practical justification.

5.5.1 Data and methods

Two biodosimetry data sets with different radiation exposure, which were previously analysed by Oliveira et al. (2016) are considered in this thesis.

The first is previously collected by Romm et al. (2013) and also analysed by Oliveira et al. (2016). The data set contains the frequency of automatically detected dicentric chromosomes, which were exposed to eight uniform doses of Cobalt-60 gamma rays. The sample size of this data set is $15,639$.

The second data set is collected by Di Giorgio et al. (2004) and also analysed by Oliveira et al. (2016) and Puig and Barquinero (2011). In the experiment, peripheral blood samples were exposed to ten doses of 1480 MeV oxygen ions and the frequency of dicentric chromosomes were recorded. This data set has sample size $8,160$. In all cases, a log-link and the quadratic model:

$$\textit{Mean number of dicentric chromosomes} \sim \textit{dose} + \textit{dose}^2 \tag{5.2}$$

is used, allowing us to obtain results that are comparable with those of Oliveira et al. (2016).

5.5.2 Results

The results for the exposure of Cobalt-60 gamma rays are given in Table 5.13. The Neyman type A gave the lowest AIC and BIC, followed by the standard compound Poisson-NB and the compound NB-Poisson models. Athough the compound NB-Poisson and the compound NB-NB models have similar AIC, the compound NB-Poisson model has a lower BIC. However, the variant models have very similar log-likelihoods as the compound models. Moreover, the difference in AIC/BIC are very small, for example the AIC of the SVA Poisson-NB model is only 7 more than that of the Neyman type A model. Hence we deduce that our proposed variant models have similar fits to the standard compound models. Diagnostic plots of distributions of residuals (see Figure 5.10) show that the variant models and the standard compound models fit equally well.

Table 5.13: *Models fitted to biodosimetry data set one.*

Models	Parameters	Log-likelihood	AIC	BIC
Neyman type A	6	$-3,738.2$	**7,488**	**7,534**
Compound Poisson-NB	7	$-3,738.1$	7,490	7,544
Compound NB-Poisson	7	$-3,739.4$	7,493	7,546
Compound NB-NB	8	$-3,738.4$	7,493	7,554
SVA Poisson-NB	7	$-3,740.5$	7,495	7,549
SVA NB-Poisson	7	$-3,741.5$	7,497	7,551
SVA NB-NB	8	$-3,740.5$	7,497	7,558
SVB Poisson-NB	7	$-3,749.4$	7,513	7,566
SVB NB-NB	8	$-3,749.4$	7,515	7,576

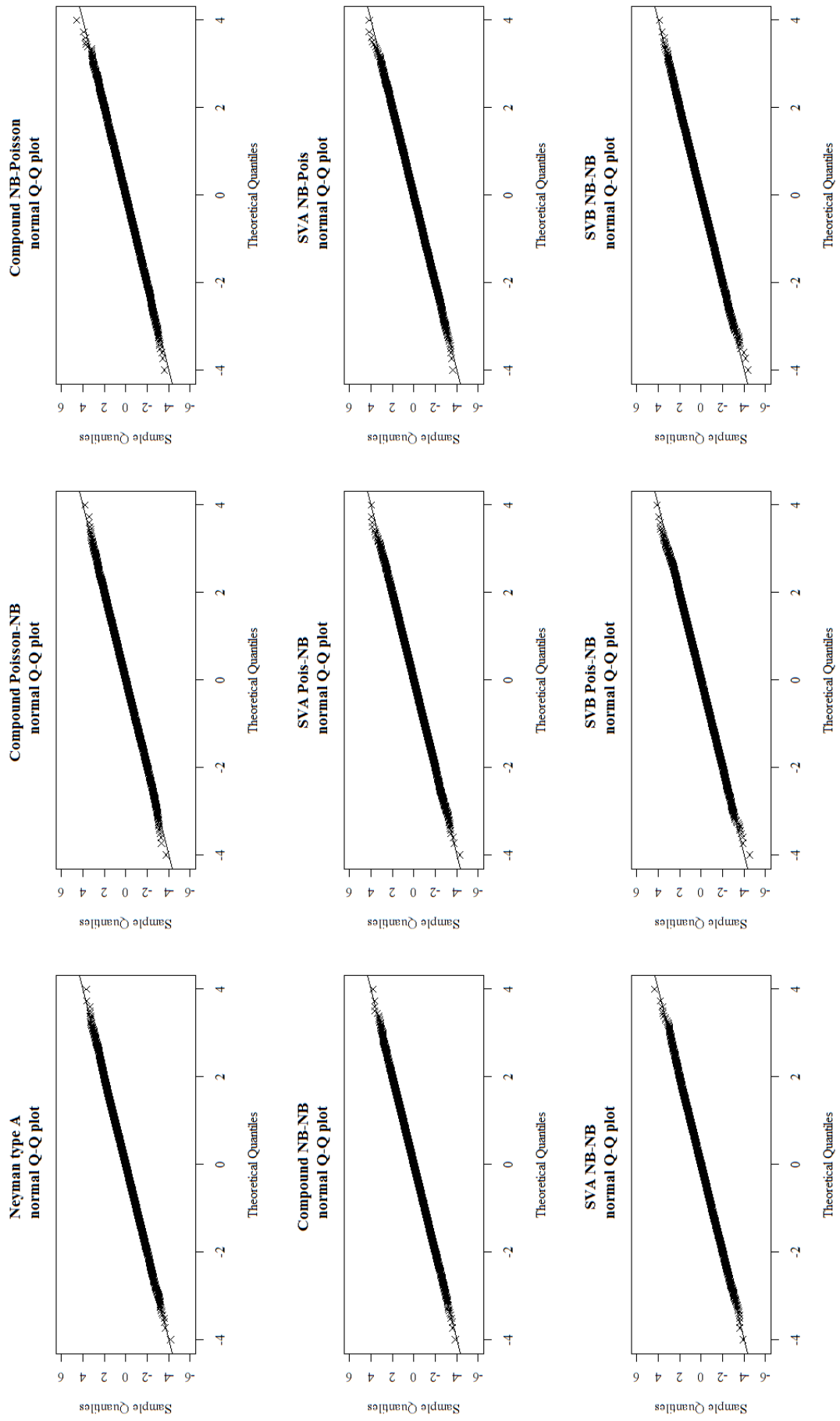

Figure 5.10: *Randomised quantile residual plots of fitted models for biodosimetry data set one.*

Table 5.14 shows that for the second data set, the Neyman type A is the superior model, followed by the compound NB-NB model and the compound NB-Poisson model. Nonetheless, our proposed variant models have similar fits to previously used models, especially in terms of log-likelihoods. For example, the log-likelihood of SVA NB-NB and SVB NB-NB is only about 6 more than that of the compound Poisson-NB model. Further examination using randomised quantile residual plots (see Figure 5.11) showed that the fits of the models, especially SVA NB-NB and SVB NB-NB, are equally good.

Table 5.14: *Models fitted to biodosimetry data set two.*

Models	Parameters	Log-likelihood	AIC	BIC
Neyman type A	6	$-2,845.2$	**5, 702**	**5, 744**
Compound Poisson-NB	7	$-2,850.9$	$5,716$	$5,765$
Compound NB-Poisson	7	$-2,847.8$	$5,710$	$5,759$
Compound NB-NB	8	$-2,843.5$	$5,703$	$5,759$
SVA Poisson-NB	7	$-2,904.5$	$5,823$	$5,872$
SVA NB-Poisson	7	$-2,904.5$	$5,823$	$5,872$
SVA NB-NB	8	$-2,856.6$	$5,729$	$5,785$
SVB Poisson-NB	7	$-2,904.5$	$5,823$	$5,872$
SVB NB-NB	8	$-2,856.6$	$5,729$	$5,785$

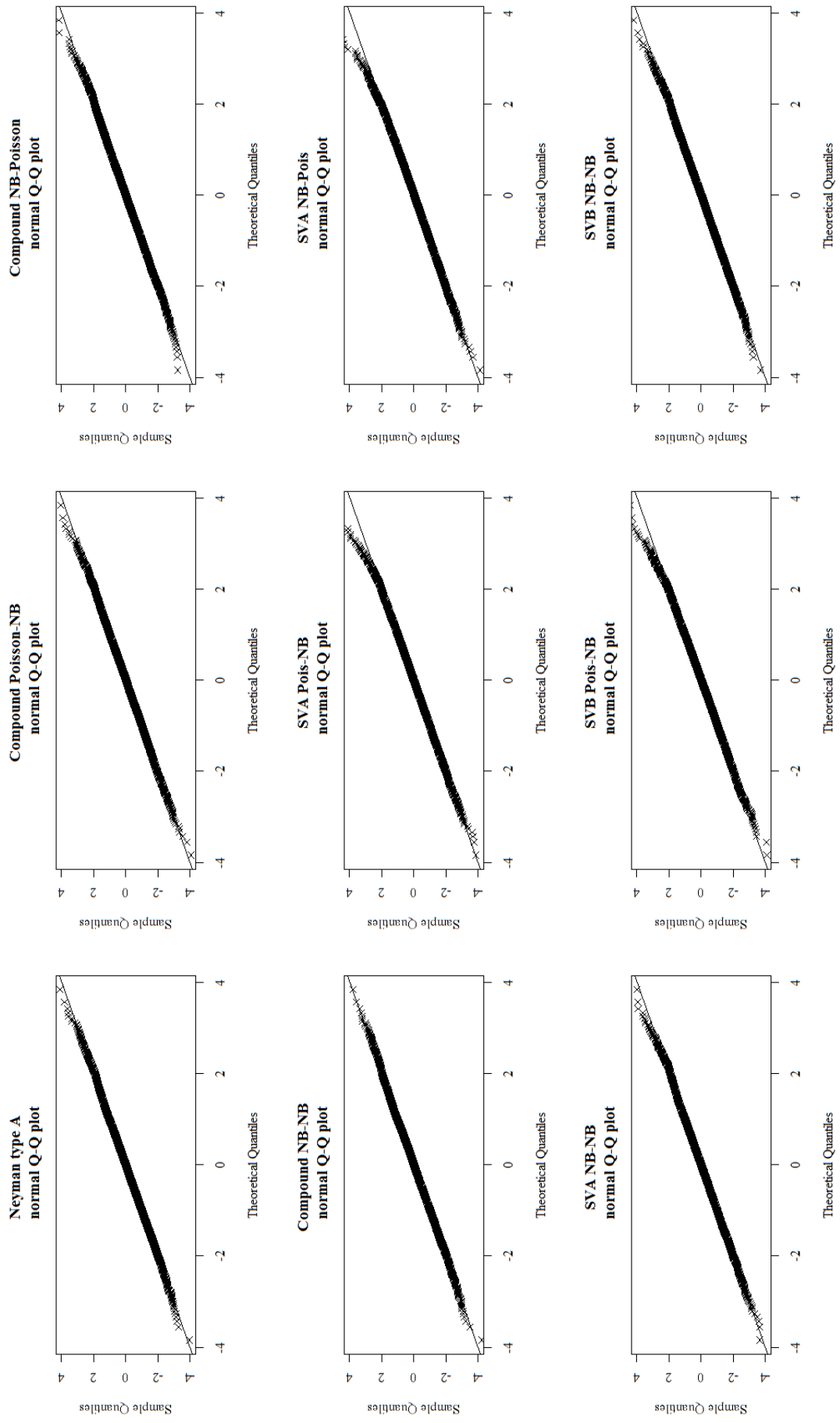

Figure 5.11: *Randomised quantile residual plots of fitted models for biodosimetry data set two.*

5.5.3 Discussion and summary

We compared the fit of standard compound models and their variants to biodosimetry data and found that their fits are quite similar especially in terms of log-likelihood. Given that there is no obvious justification for the use of these variants, as it is difficult to interpret the number of abberated chromosomes as two generations, it is unsurprising that these variants are not selected by the model selection criteria, whilst the standard compound models, which have appropriate interpretations, are favoured instead. Thus, it is arguable that even if the AIC/BIC of the variant models were superior, the standard compound models should still be used.

Chapter 6

Christmas tree plots for model validation

In the previous chapter, randomised quantile residual plots were used for model checking. In this chapter, we investigate the use of an alternative method for model validation to see if similar results will be obtained. Wilson and Einbeck (2015) have proposed a new test to check for number inflation or deflation relative to a count model. For example, if data are fitted with a Poisson model, the test examines the expected number of $0, 1, 2, \ldots$ relative to the Poisson model and compares these with the data. A diagnostic plot, referred as a "Christmas tree plot" was also introduced to illustrate the results from the test so that the suitability of a model may be assessed diagrammatically. Initial research by Wilson and Einbeck (2015, 2016) focused mainly on testing the null hypothesis of a Poisson model against the alternative hypothesis of a zero-modified Poisson model. Einbeck and Wilson (2016) extended this by providing guidelines to assess model fits for any count regression models. In this chapter, we investigate the use of Christmas tree plots by extending this method to our proposed variant models applied in citation analysis.

6.1 Background

Let $Y = \{y_1, y_2, \ldots, y_n\}$ be n independent random observations, derived from a count distribution, $F(\mu_i, \Theta)$, where Θ is a vector of covariates, and $\mu_i = E(y_i|\theta_i)$. Say $c \in [0, max(Y)]$, and suppose we wish to investigate if the number of observed c, $N(c)$, in Y is consistent with a fitted distribution F.

Let $p_i(c) = P(Y = c|\hat{\mu}_i, \hat{\theta}_i)$, and X_c be a Bernoulli random variable with

parameter $p_i(c)$ where:

$$X_c = \begin{cases} 1 & \text{if } Y = c \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

In the presence of covariates, that is, when μ_i depends on covariates, $p_i(c)$ varies, X_c is a Bernoulli random variable with varying p_i, then the sum, $N(c) = X_1 + X_2 + \ldots + X_c$ is a Poisson-Binomial random variable with parameters $p_1(c), p_2(c), \ldots p_n(c)$ and pmf:

$$P(N(c) = k) = \left\{ \prod_{i=1}^{n} (1 - p_i(c)) \right\} \sum_{i_1 < \ldots < i_k} w_{i1} \ldots w_{ik} \tag{6.2}$$

where $w_i = \dfrac{p_i(c)}{1 - p_i(c)}$, for $i = 1, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, \ldots i_k$ from $\{1, \ldots, n\}$ (Chen and Liu, 1997). The R package *poibin* (Hong, 2013) may be used for the computation of the Poisson-Binomial distribution. Note that the Poisson-Binomial distribution is different from a compound Poisson Binomial distribution. In the absence of covariates, all the μ_i are equal, $N(c)$ is the sum of n independent X_i Bernoulli random variables with equal $p_i(c)$, and $N(c)$ is therefore a binomial random variable.

This test enables the computation of upper and lower limits for the expected values of $N(c)$, known as fluctuation intervals, at some significance level α. The fluctuation interval is a similar concept to the confidence interval. Whilst confidence intervals are defined for an estimated statistical parameter, this is not the case for fluctuation intervals. For instance, a $(1 - \alpha) \times 100\%$ fluctuation interval for $N(k)$ has lower and upper limits, denoted $l_\alpha(c)$ and $u_\alpha(c)$ respectively. If approximately $(1 - \alpha) \times 100\%$ of $N(c)$ values are within the limits, then it is consistent with the fitted distribution F.

The techniques of this test may be illustrated in diagnostic plots known as "Christmas tree plots" for visualisation purposes. A median adjustment has been proposed to obtain a clearer plot. Following Einbeck and Wilson (2016), these plots are constructed for a chosen interval, $C = [c_a, c_b]$, by:

(i) Fitting the model, $F(\mu_i, \theta_i)$ to the data, Y and obtaining the estimates, $\hat{\mu}_i$ and $\hat{\theta}$ from the fitted model.

(ii) Obtaining estimates of $\hat{p}_i(c)$ for c in $c_a, \ldots, c_b$. Using a Poisson-Binomial distribution under count data model F, estimate the median, $m(c) = med(N(c))$, lower, $l_\alpha(c)$ and upper, $u_\alpha(c)$ limits of the $(1 - \alpha) \times 100\%$ fluctuation interval for $N(c)$.

(iii) Adjusting the limits by computing the median adjusted bounds, $\underline{b}_\alpha(c) = l_\alpha(c) - m(c)$ and $\bar{b}_\alpha(c) = u_\alpha(c) - m(c)$.

(iv) Plotting the functions $\underline{b}_\alpha(c)$ and $\bar{b}_\alpha(c)$ against c. Adding the adjusted observed counts, $A(c) = N(c) - m(c)$ of the observed data Y to the same plot to obtain reasonable comparisons.

Other possible further adjustments to the response variable, such as applying the natural log or using the square root upon median adjustment have also been investigated but these did not improve the clarity of the plots. Thus, only the median adjustment is used here.

6.1.1 Example

We illustrate the test using the classical example of horse kicks data (von Bortkiewicz, 1898), which contains the number of deaths of soldiers in the Prussian army from horse kicks. The soldiers were from 10 army corps and the deaths were observed over 20 years. If the Poisson model is fitted to the data and 90% fluctuation intervals are computed, then the observed number, lower and upper limits of the fluctuation intervals relative to the Poisson model are given in Table 6.1.

Table 6.1: *Horse kick data with lower and upper limits for* 90% *fluctuation intervals.*

c	$N(c)$	$l_{0.10}$	$u_{0.10}$	$m(c)$	$\underline{b}_{0.10}(c)$	$\bar{b}_{0.10}(c)$
0	109	97	120	109	−12	11
1	65	55	77	66	−11	11
2	22	13	27	20	−7	7
3	3	1	8	4	−3	4
4	1	0	2	0	0	2
5	0	0	1	0	0	1

The results in Table 6.1 may be illustrated using Christmas tree plots, as shown in Figures 6.1 and 6.2. Although the median adjustment is not used in Figure 6.1, it is still clear that the observed counts are within the upper and lower boundaries. Figure 6.2 shows the Christmas tree plot when a median adjustment is applied to the boundaries and the counts. This is useful as it will give a clearer illustration in cases when the boundaries are very close.

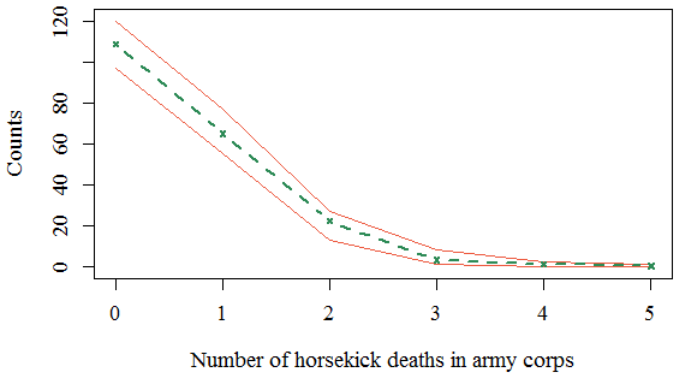

Figure 6.1: *A Christmas tree plot for the horsekick data, relative to a Poisson model. The orange lines are the boundaries while the green crosses are the observed counts.*

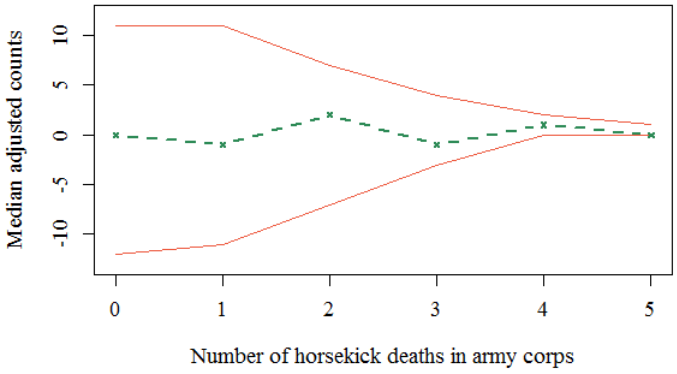

Figure 6.2: *A Christmas tree plot for the horsekick data using median adjusted counts, relative to a Poisson model. The orange lines are the adjusted boundaries while the green crosses are the median adjusted counts.*

Einbeck and Wilson (2016) have provided a comprehensive guideline to generalise this test to any count distribution, and applied this to biodosimetry data which consist of small data values of 0 up to 7. We further extend the application of this test using citation data, which spans a much larger range of data values, relative to proposed distributions in this thesis.

6.2 Christmas tree plots for simulated SVA and SVB data

The use of diagnostic plots are first investigated using simulated data from SVA and SVB distributions. This is carried out using randomly selected fixed pa-

rameter values for each distribution. In each case, data points are simulated and refitted using the same distribution without incorporating any covariates. For example, the diagnostic plot for simulated data using the SVA Poisson-NB$(3, 2, 1)$ distribution is given in Figure 6.3. As expected, the observations in Figure 6.3 are consistent with their probabilities. The median adjusted counts of the first few numbers are zeros because the probabilities of those points are also close to zero. The probabilities then increase for data values up to about 25, and then decrease. Since the fitted model is also the simulation model, it is unsurprising that the observed counts are scattered largely within the fluctuation intervals, indicating a good fit.

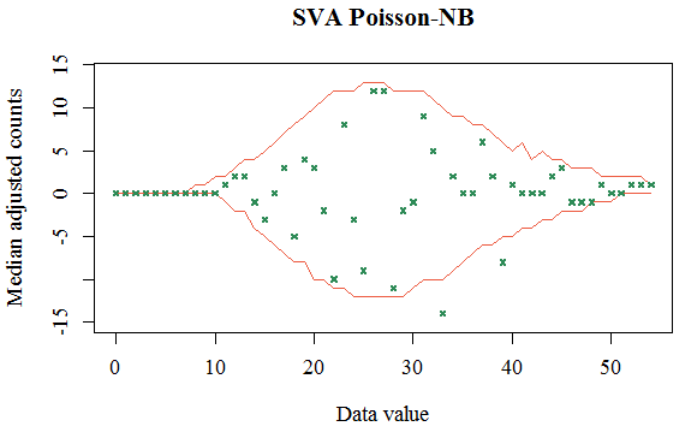

Figure 6.3: *A Christmas tree plot for* 1000 *data simulated from a SVA Poisson-NB*$(3, 2, 1)$ *distribution. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

Similar plots are obtained for data simulated from SVA NB-Poisson and SVA NB-NB distributions (see Figures 6.4 and 6.5), indicating suitable fits. In both plots, approximately 90% of the median adjusted counts are within the boundaries, indicating that the models are adequate.

One puzzling feature of the plots is the presence of unexpected spikes in the fluctuation bounds. This may be a consequence of estimating quantiles of discrete data. Adoption of methods proposed by Ma et al. (2011) to formulate quantiles of discrete distributions may remove this feature but this is not considered here as it will further complicate the testing procedure.

Figures 6.6 and 6.7 show the Christmas tree plots when data are simulated and refitted using the proposed SVB distributions. In both diagrams, approximately 90% of the observed median adjusted counts lie within the boundaries, indicating a good fit.

Overall, this section gives evidence that if data are simulated from SVA and SVB distributions and refitted with the generating model, then the Christmas

tree plots indicate that the models are adequate, because approximately 90% of the points are within the boundaries.

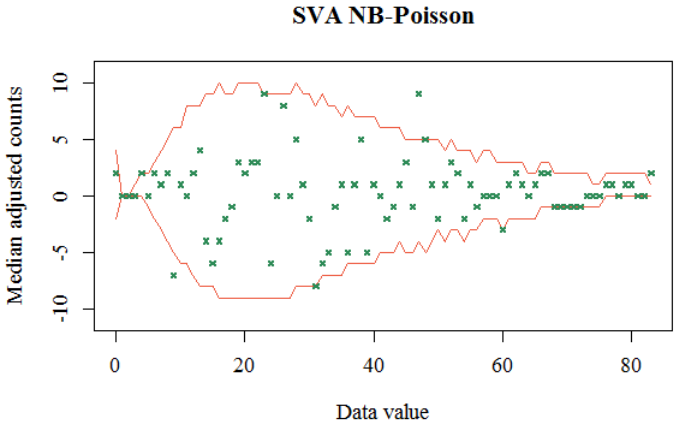

Figure 6.4: *A Christmas tree plot for* 1000 *data simulated from a SVA NB-Poisson*$(3, 1, 2)$ *distribution. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

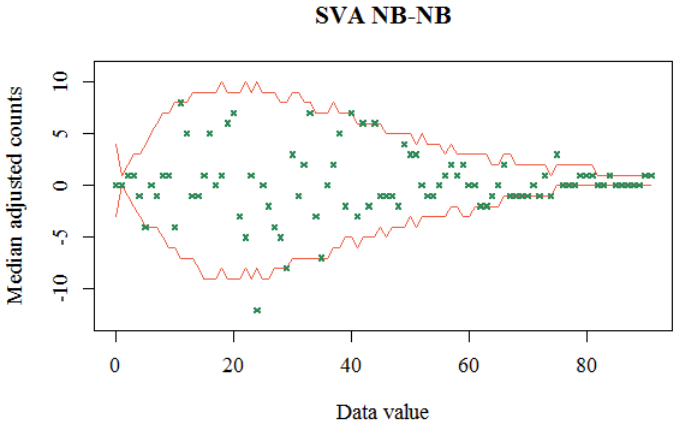

Figure 6.5: *A Christmas tree plot for* 1000 *data simulated from a SVA NB-NB*$(3, 1, 2, 1)$ *distribution. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

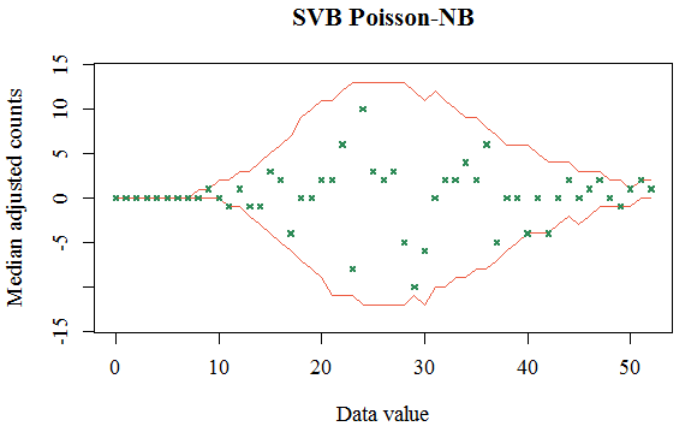

Figure 6.6: *A Christmas tree plot for* 1000 *data simulated from a SVB Poisson-NB*$(3, 2, 1)$ *distribution. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

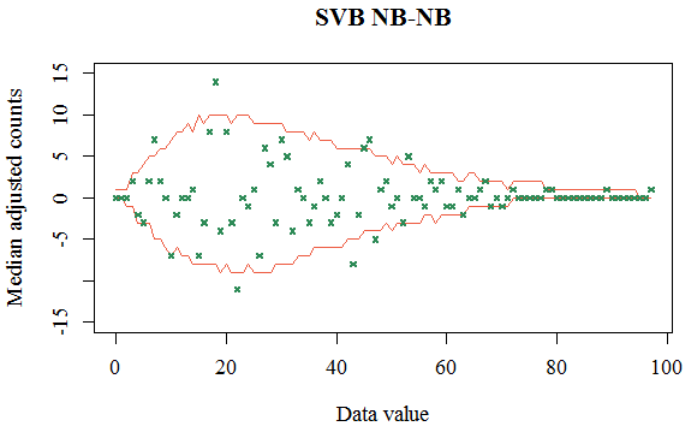

Figure 6.7: *A Christmas tree plot for* 1000 *data simulated from a SVB NB-NB*$(3, 1, 2, 1)$ *distribution. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

6.3 Christmas tree plots for citation data with no covariates

This section applies the Christmas tree plots to the citation data of Section 5.3. The median adjustment is used for all plots to produce clearer diagrams and the plots for *Tourism* are discussed in detail to allow comparison with the results made from the randomised quantile residual plots in Section 5.3.2. The Christmas tree plots for the other subjects are presented in Appendix G.

6.3.1 Tourism

Using the procedures discussed in Section 6.1 and in Einbeck and Wilson (2016), a Christmas tree plot is obtained. Figure 6.8 shows the Christmas tree plot when a negative binomial model is fitted to Tourism data.

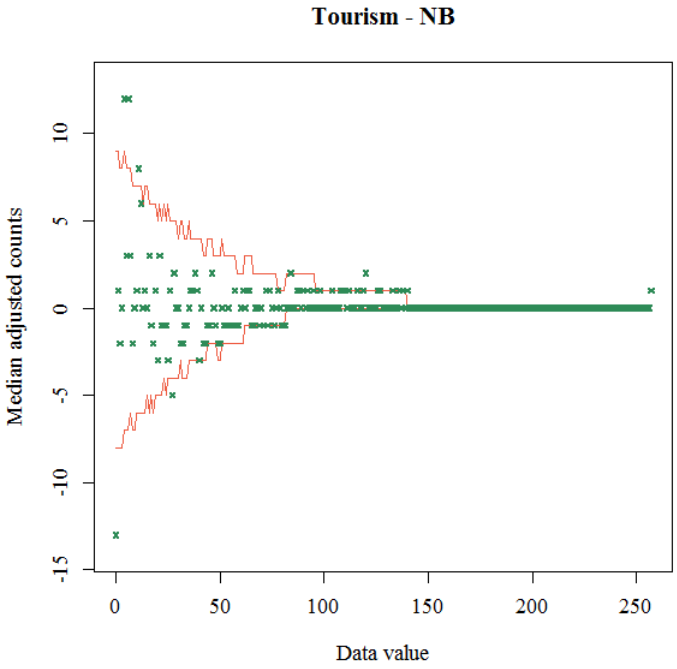

Figure 6.8: *A Christmas tree plot for Tourism when fitted with the negative binomial model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

In Figure 6.8, a few points are outside of the boundaries. For example, when $c = 0, A(0) = -13$, but this is less than the lower boundary, as $\underline{b}_{0.1}(0) = -8$. The other cases when the median adjusted counts are outside the boundaries of the 90% fluctuation intervals are given in Table 6.2. Nonetheless, this is still acceptable as approximately 90% of the adjusted counts are still within the boundaries.

Table 6.2: *Cases when Tourism citation counts are outside the adjusted boundaries for 90% fluctuation intervals when fitted with the negative binomial model.*

c	$N(c)$	$l_{0.10}$	$u_{0.10}$	$m(c)$	$A(c)$	$\underline{b}_{0.10}(c)$	$\bar{b}_{0.10}(c)$
0	15	20	37	28	−13	−8	9
4	34	15	31	22	12	−7	9
6	32	13	28	20	12	−7	8
11	24	10	23	16	8	−6	7
27	3	4	13	8	−5	−4	5
120	2	0	1	0	2	0	1
140	1	0	0	0	1	0	0
257	1	0	0	0	1	0	0

At about $c = 140$, the lines indicating the upper and lower boundaries merge. This is because for $c = 140$, $P(c \geq 140) < 0.10$. Hence, based on the Poisson-Binomial distribution, $P(N(c) = 0 \,|\, c \geq 140) > 0.90$. Thus, for any $k \leq 90$, the k^{th} quantile is 0. We regard the point at which the upper and lower boundaries intersect as a threshold. Consequently, $c = 140$ is used as the threshold and this corresponds to the 99.8^{th} quantile. In the observed data, a data value $c = 140$, which lies on the threshold, was observed. The point $c = 257$ was also observed that is greater than the threshold and thus this point is inconsistent with the negative binomial model (as illustrated in Figure 6.8). This may occur because there exist covariates which boost the citation counts of this particular article, but are omitted in the model. Thus this point may escalate the mean or largely influence the dispersion of the negative binomial model.

For the fitted discretised lognormal, SVA and SVB models, the number of observations outside the boundaries, threshold and number of points beyond the threshold are recorded (see Table 6.3). Although the threshold values differ for different models, only one observed count ($c = 257$) lies beyond the thresholds for all models considered here for Tourism.

Table 6.3: *Results for Tourism.*

Models	Number of observations outside the boundaries	Threshold	Number of observations beyond the threshold
Discretised lognormal	2	$c = 192$	1
SVA Poisson-NB	5	$c = 162$	1
SVA NB-Poisson	8	$c = 140$	1
SVA NB-NB	4	$c = 183$	1
SVB Poisson-NB	3	$c = 151$	1
SVB NB-NB	2	$c = 170$	1

Prior to fitting the discretised lognormal model, one is added to all citation counts in Tourism. The Christmas tree plot relative to the discretised lognormal model in Figure 6.9 indicates a good fit.

If the SVA Poisson-NB model is fitted to Tourism data, five points lie outside the boundaries. The details of these 5 points are given in Table 6.4. The associated Christmas tree plot in Figure 6.10 indicates that the SVA Poisson-NB model is adequate.

Table 6.4: *Cases when Tourism citation counts are outside the adjusted boundaries for* 90% *fluctuation interval when fitted with the SVA Poisson-NB model.*

c	$N(c)$	$l_{0.10}$	$u_{0.10}$	$m(c)$	$A(c)$	$\underline{b}_{0.10}(c)$	$\bar{b}_{0.10}(c)$
0	15	17	32	24	−9	−7	8
1	27	6	16	10	17	−4	6
5	24	25	44	34	−10	−9	10
120	2	0	1	0	2	0	1
257	1	0	0	0	1	0	0

When the SVA NB-Poisson model is fitted to Tourism data, 8 data points ($c = 0, 4, 6, 11, 27, 120, 140, 257$) lie outside the boundaries. When the SVA NB-NB model is fitted to Tourism data, only 4 data points ($c = 0, 4, 120, 257$) lie outside the boundaries. Both diagrams in Figures 6.11 and 6.12 indicate that the respective models are suitable.

The Christmas tree plots for Tourism when SVB Poisson-NB and SVB NB-NB models are fitted indicate that both models are adequate (see Figures 6.13

and 6.14).

In all cases, approximately 90% of the median adjusted counts lie within the upper and lower boundaries, indicating that overall, the negative binomial, discretised lognormal and all the SVA and SVB models investigated are suitable. The results obtained here are consistent with those from the randomised quantile residual plots in Section 5.3.2.

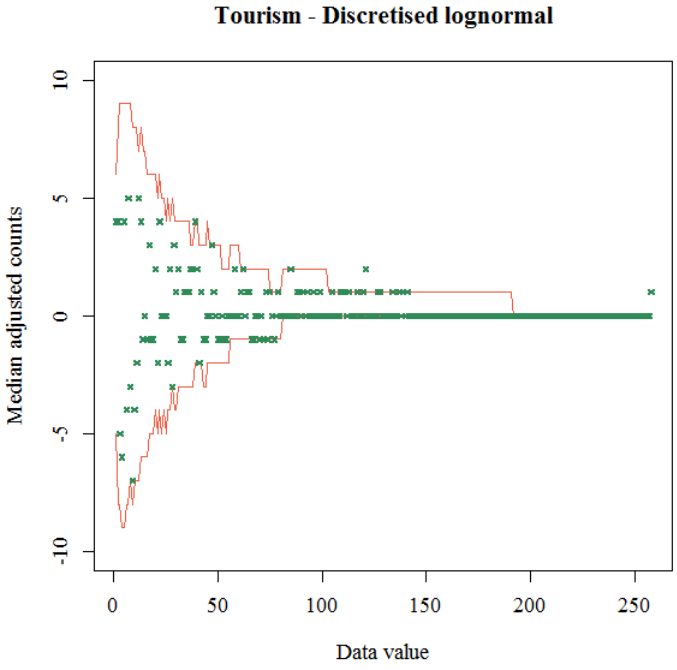

Figure 6.9: *A Christmas tree plot for Tourism when fitted with discretised lognormal model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

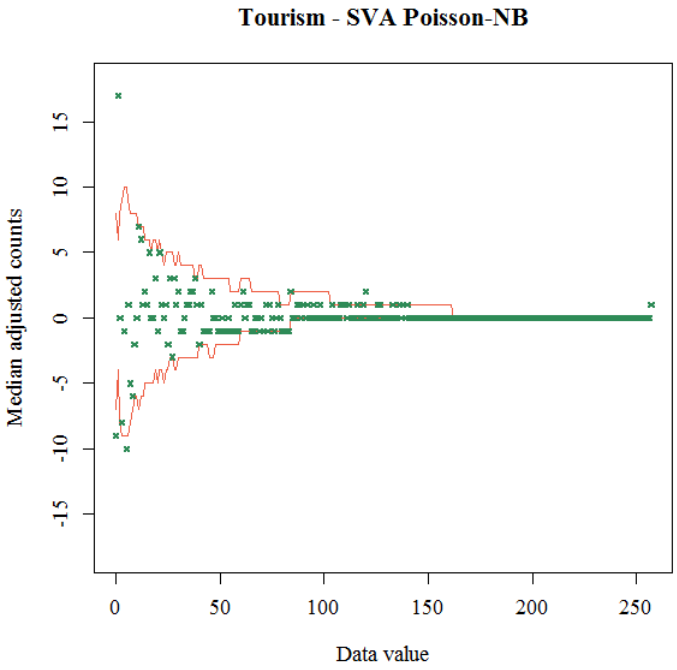

Figure 6.10: *A Christmas tree plot for Tourism when fitted with SVA Poisson-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

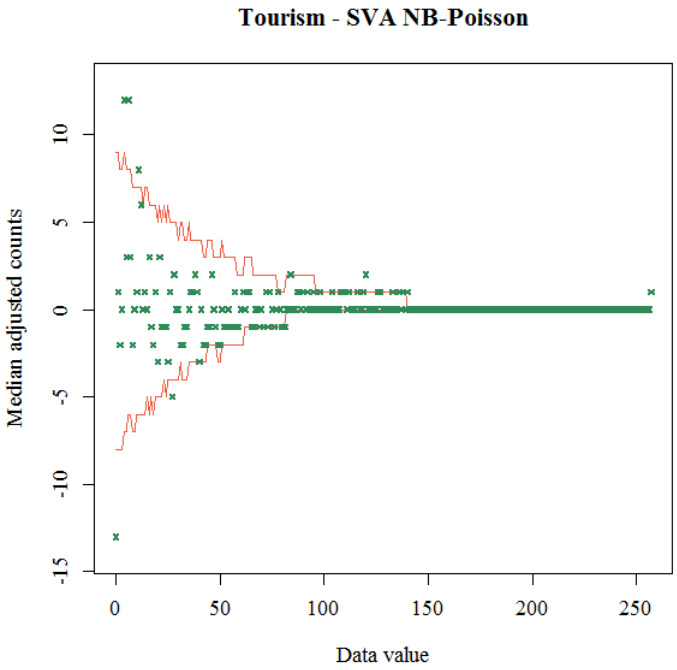

Figure 6.11: *A Christmas tree plot for Tourism when fitted with the SVA NB-Poisson model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

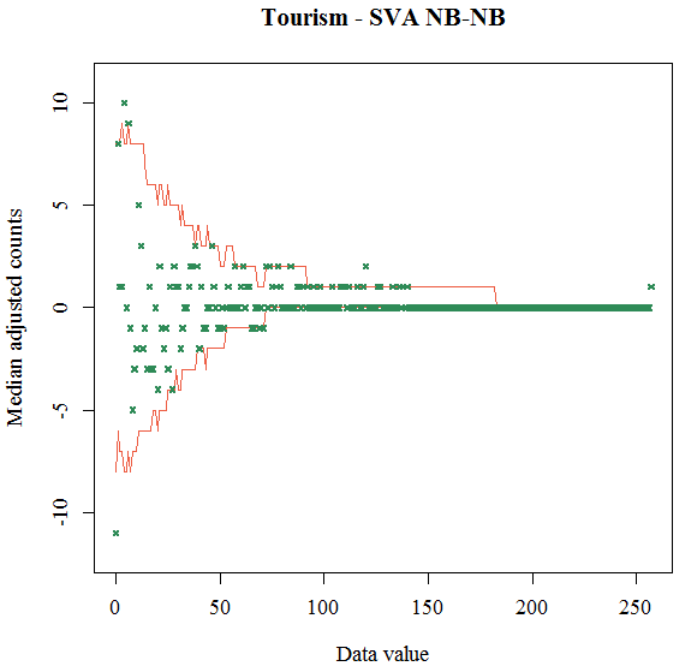

Figure 6.12: *A Christmas tree plot for Tourism when fitted with the SVA NB-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

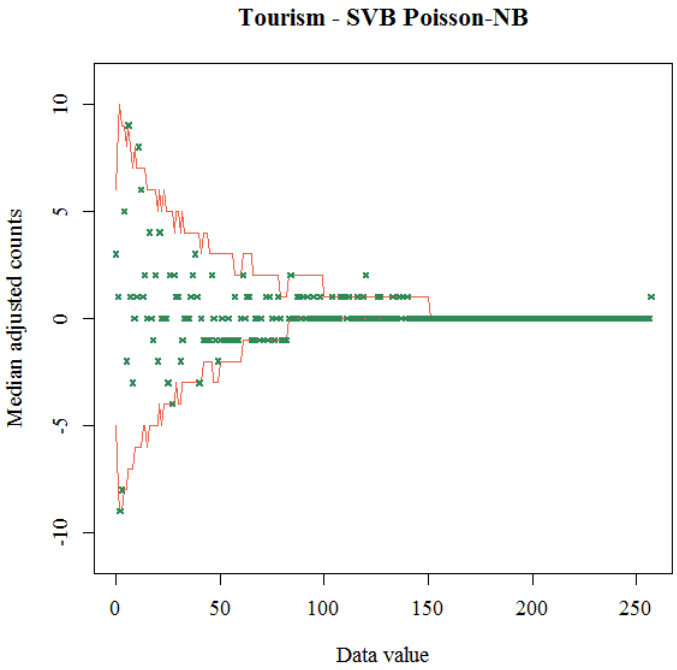

Figure 6.13: *A Christmas tree plot for Tourism when fitted with the SVB Poisson-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

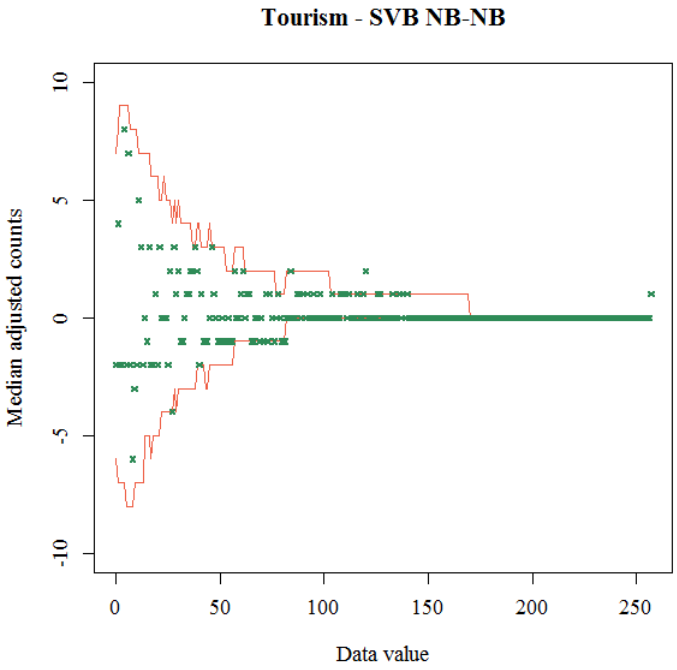

Figure 6.14: *A Christmas tree plot for Tourism when fitted with the SVB NB-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

6.4 Christmas tree plots for citation data with covariates

This section applies diagnostic plots to the citation analysis in Section 5.4. The plots for *Applied Mathematics* and *Aquatic Science* are discussed in detail so that comparisons can be made with the results obtained from the randomised quantile residual plots in Section 5.4.2. The Christmas tree plots for the other subjects are given in Appendix H.

6.4.1 Applied Mathematics

When the negative binomial model is fitted to Applied Mathematics citation data, the test shows that 30 points are outside the boundaries, of which 16 are between 0 and 34. This exceeds the expected variation by the model, indicating that the negative binomial model may be unsuitable. This is also illustrated in Figure 6.16.

The results obtained for Applied Mathematics when fitted with SVA and SVB models are given in Table 6.5, and their respective Christmas tree plots are given in Figures 6.17 to 6.21.

Table 6.5: *Results for Applied Mathematics.*

Models	Number of observations outside the boundaries	Threshold	Number of observations beyond the threshold
SVA Poisson-NB	20	$c = 314$	3
SVA NB-Poisson	30	$c = 184$	9
SVA NB-NB	33	$c = 326$	3
SVB Poisson-NB	30	$c = 183$	9
SVB NB-NB	24	$c = 230$	7

If the SVA Poisson-NB model is fitted to Applied Mathematics data, then the results show that 20 points lie outside the boundaries. Amongst these points, three ($c = 384, 651, 743$) are far beyond the threshold of $c = 314$. For the SVA NB-Poisson model, the results show that 30 points are outside the boundaries and 16 of these are in the range of 0 to 34. This indicates that the model did not fit the data well, especially for small data values. Similar results are obtained for the SVA NB-NB model. A total of 33 points are outside the boundaries and 25 of these points are between 0 and 48. The Christmas tree plots for the fitted SVA models indicate that the SVA models are unsuitable for Applied Mathematics (see Figures 6.17 to 6.19).

When the SVB Poisson-NB model is fitted to Applied Mathematics, the results show that 30 points lie outside the boundaries. The first 16 of these are between 0 and 34, indicating that this model is unsuitable. This is also illustrated in the Christmas tree plot in Figure 6.20. For the SVB NB-NB model, 14 out of the 24 points which are outside the fluctuation bounds are between 0 and 35. The plot in Figure 6.21 shows that the SVB NB-NB model is unsuitable for Applied Mathematics.

Overall, the Christmas tree plots for Applied Mathematics suggest that the negative binomial, SVA and SVB models fitted may be unsuitable, especially for small data values ($c < 48$). Whilst the presence of large positive residuals in the randomised quantile residual plots are consistent with the Christmas tree plots due to the presence of observed counts beyond the threshold, the randomised quantile residual plots seem to show no issue with small data values (see Figures 5.6 and 5.7). In fact, this discrepancy is false. If for example, the plot for the negative binomial model is enlarged, as shown in Figure 6.15, then it is clear that the points do not lie on the line. More specifically, majority of the points are

above the line when $z < -0.3$, which is approximately the 38^{th} or lower percentile and represents $c \leq 2$ in this case, indicating that the model underestimates the number of observations for $c \leq 2$.

Table 6.6: *The first three observations for Applied Mathematics citation counts when fitted with the negative binomial model, with their* 90% *fluctuation intervals.*

c	$N(c)$	$l_{0.10}$	$u_{0.10}$	$m(c)$	$A(c)$	$\underline{b}_{0.10}(c)$	$\bar{b}_{0.10}(c)$
0	1229	1336	1444	1390	−161	−54	54
1	793	680	764	722	71	−42	42
2	629	488	560	524	105	−36	36

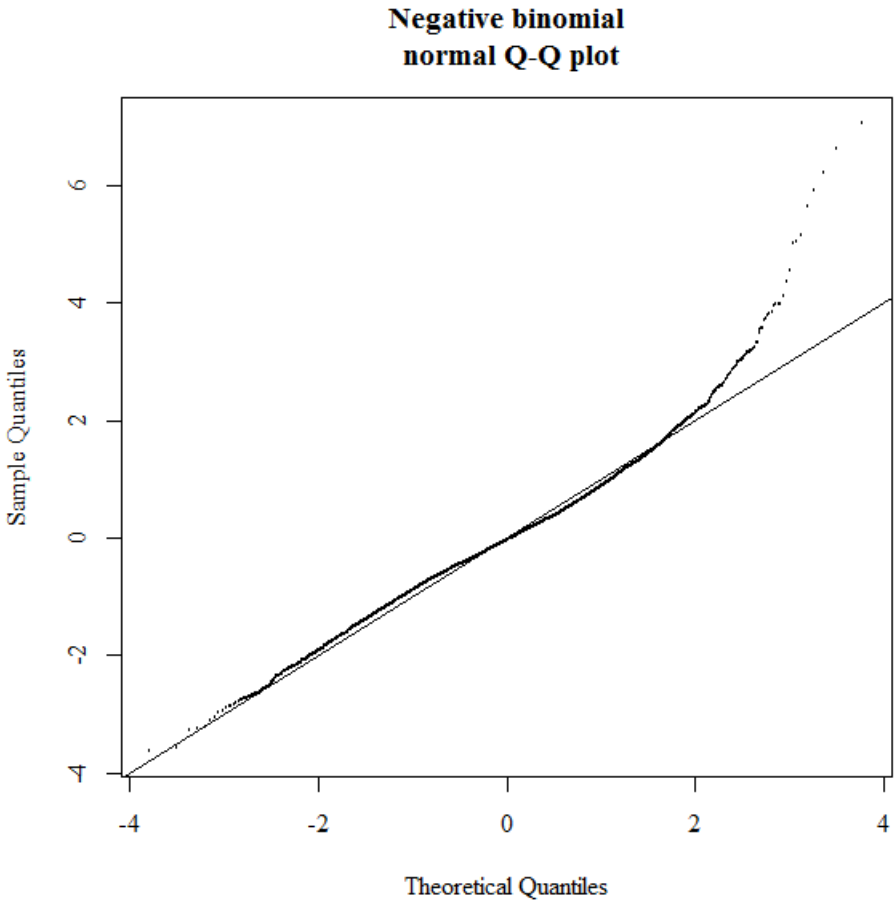

Figure 6.15: *An enlarged randomised quantile residual plot for Applied Mathematics when fitted with the negative binomial model.*

Here, a total of 6411 citation counts are observed, of which 19.2% are zeros and 12.4% are ones. However, under the fitted negative binomial model, we expect 21.7% zeros (since $1390/6411 = 0.217$) and 11.3% ones (since $722/6411 = 0.113$). The respective $N(c)$ and $m(c)$ for the first three data values are given in Table 6.6.

Using $c = 0$ as an example, $P(Z < -0.87) = 0.192$ but $P(Z < -0.78) = 0.217$, thus the difference in quantiles are very small (0.09). Cumulatively, we observe $P(c \leq 1) = 0.315$, but under the negative binomial model, $P(c \leq 1) = 0.329$. As the cumulative probabilities for $c \leq 1$ are very close, this difference is not apparent in the randomised quantile residual plot.

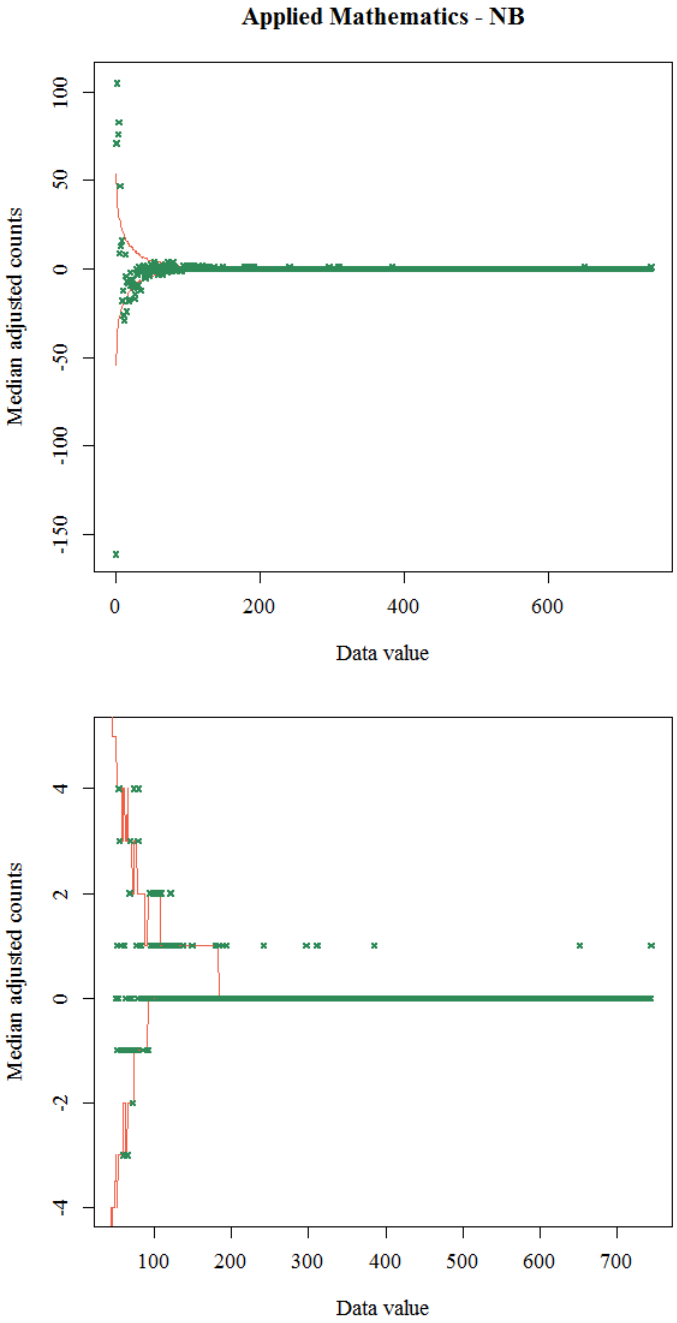

Figure 6.16: *A Christmas tree plot for Applied Mathematics when fitted with the negative binomial model (top). The bottom plot magnifies the top plot for data values greater than* 50*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

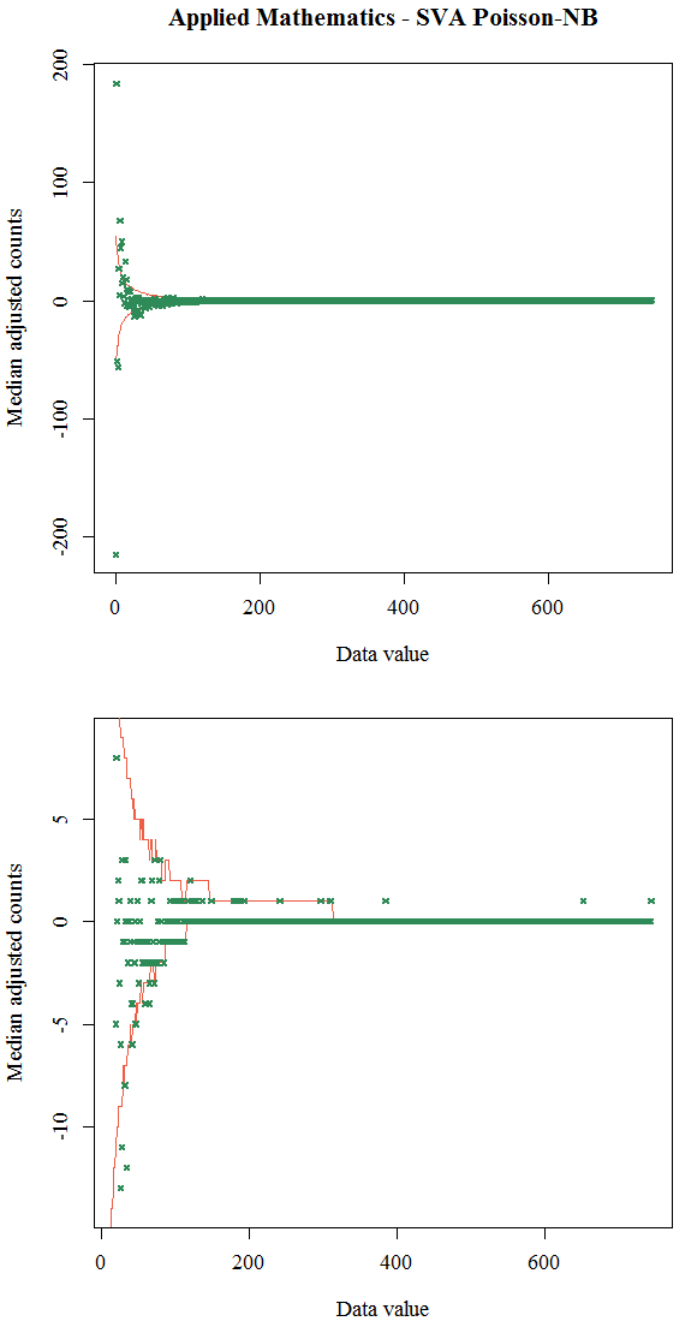

Figure 6.17: *A Christmas tree plot for Applied Mathematics when fitted with the SVA Poisson-NB model (top). The bottom plot magnifies the top plot for data values greater than* 20*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

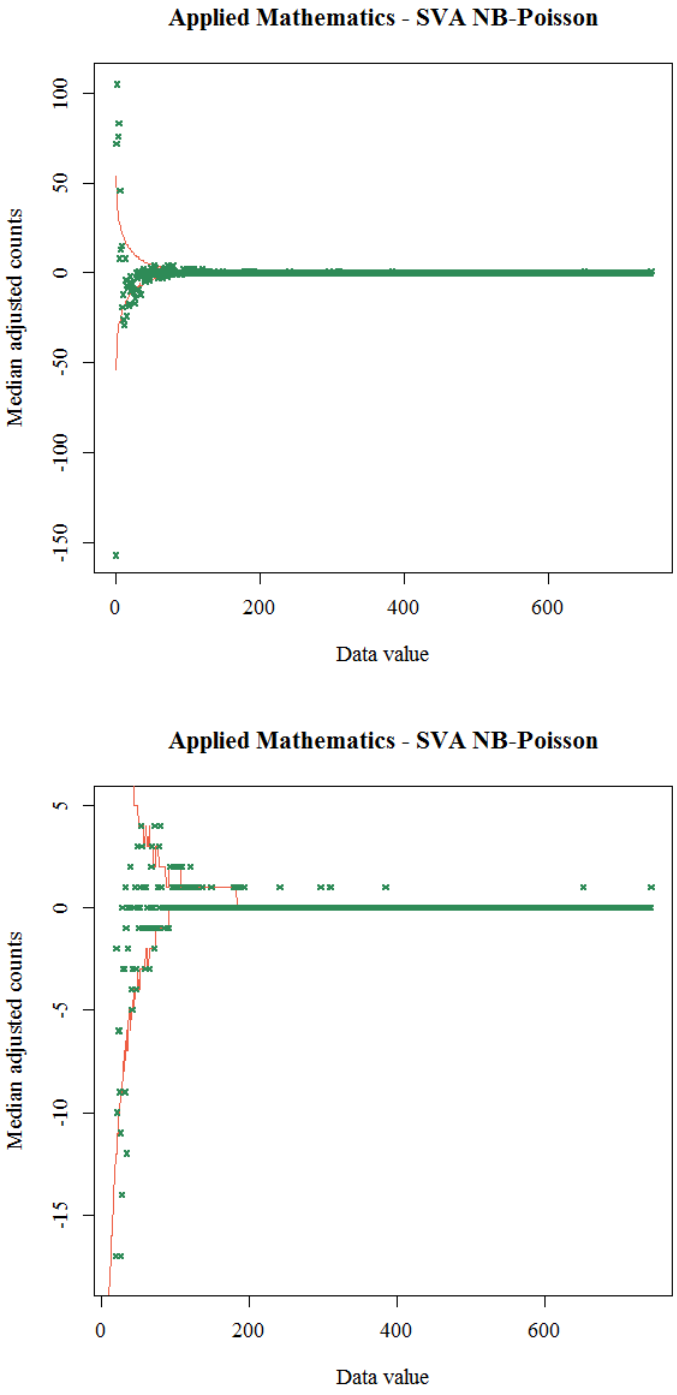

Figure 6.18: *A Christmas tree plot for Applied Mathematics when fitted with the SVA NB-Poisson model (top). The bottom plot magnifies the top plot for data values greater than* 20*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

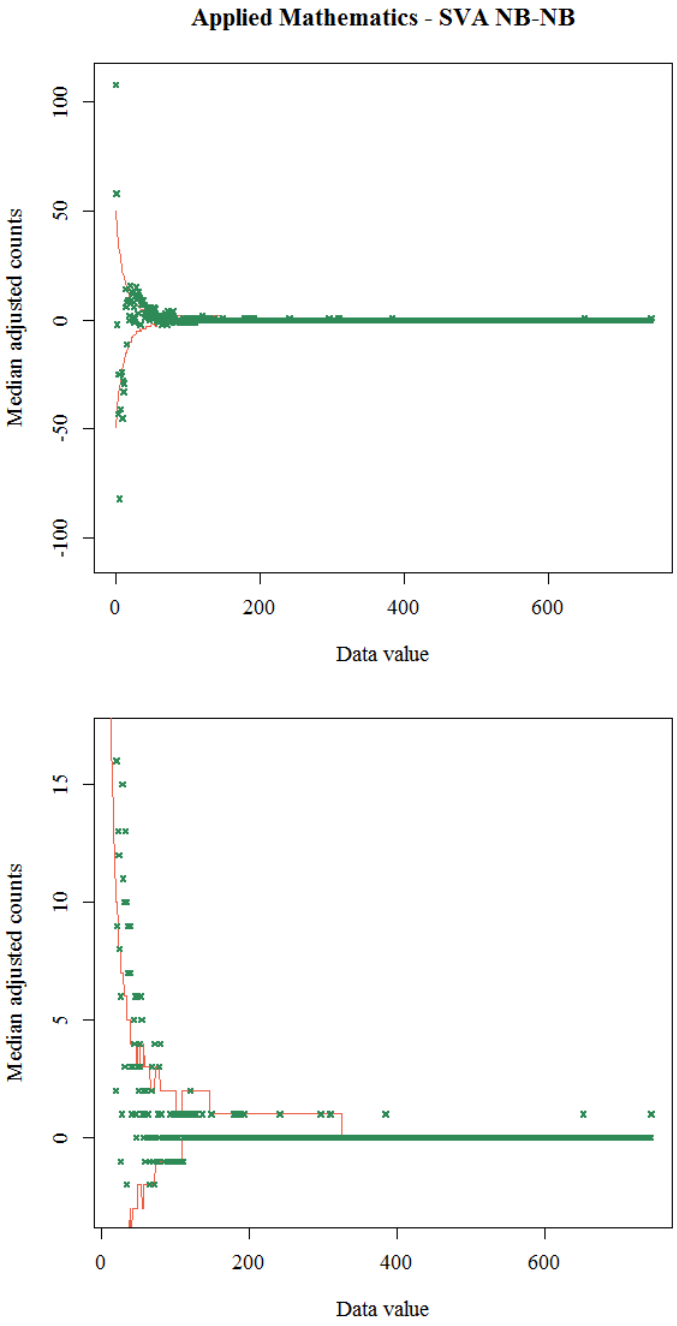

Figure 6.19: *A Christmas tree plot for Applied Mathematics when fitted with the SVA NB-NB model (top). The bottom plot magnifies the top plot for data values greater than* 20*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

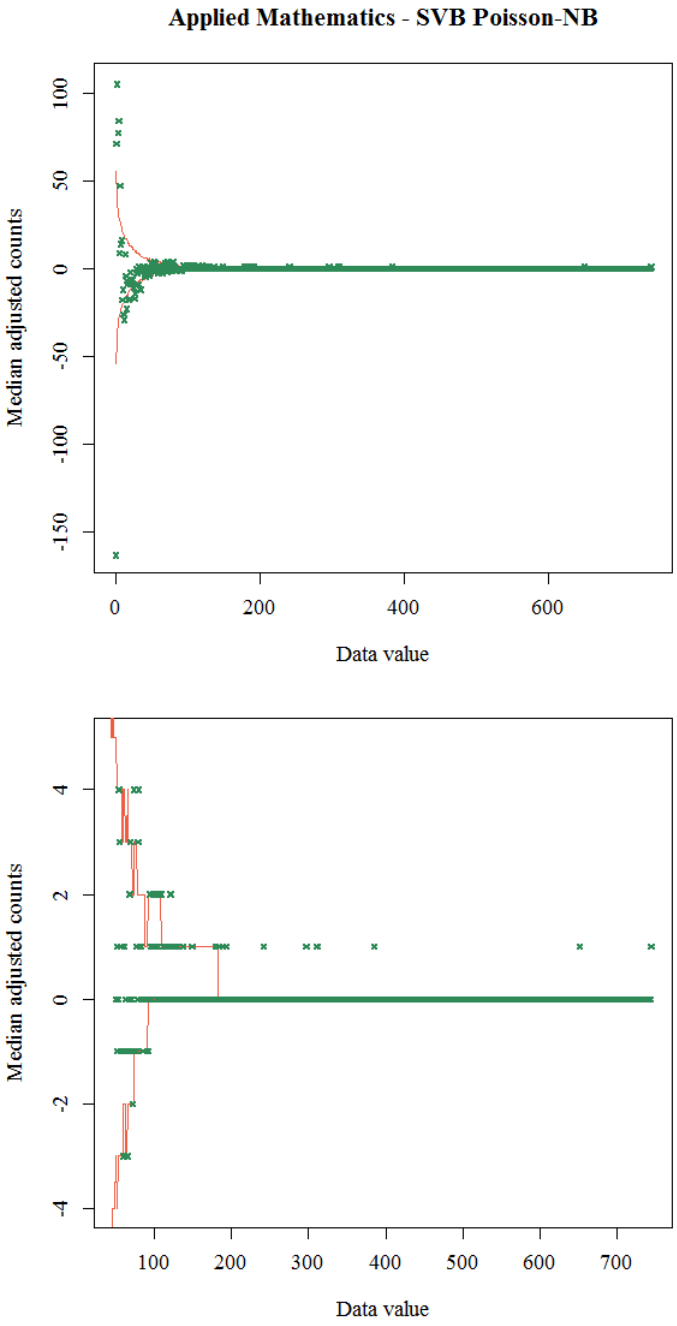

Figure 6.20: *A Christmas tree plot for Applied Mathematics when fitted with the SVB Poisson-NB model (top). The bottom plot magnifies the top plot for data values greater than* 50*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

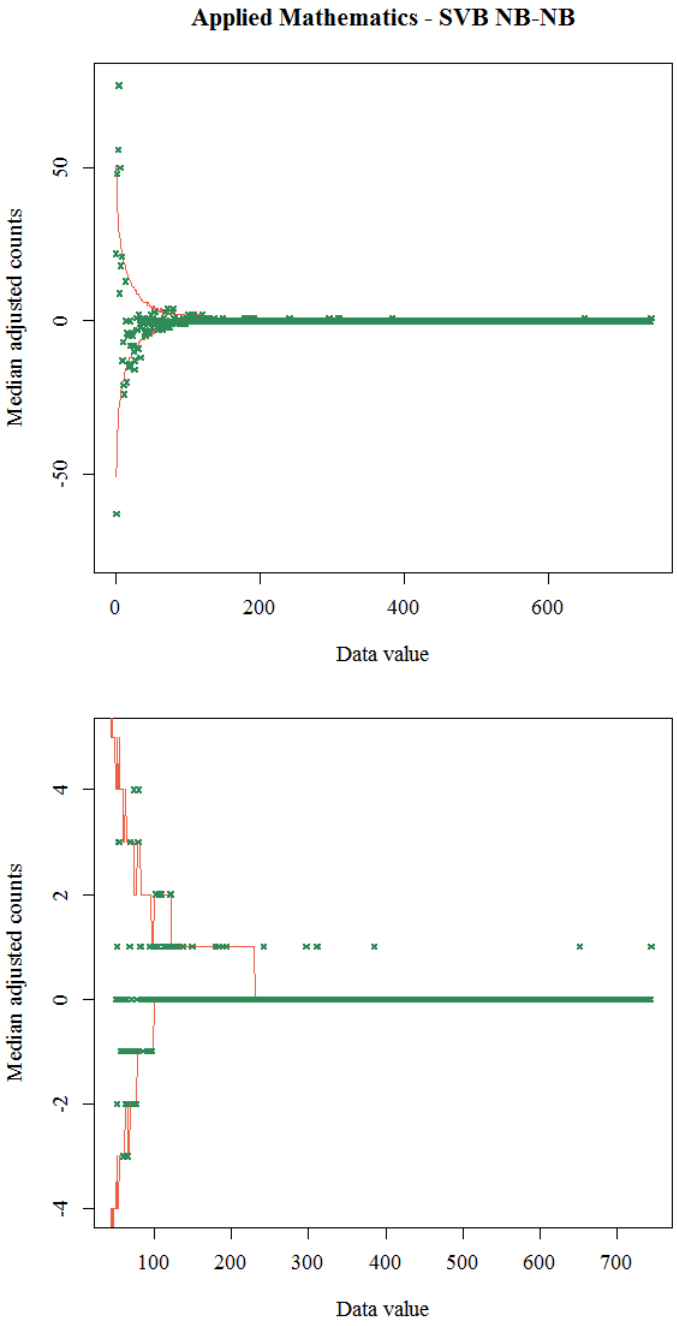

Figure 6.21: *A Christmas tree plot for Applied Mathematics when fitted with the SVB NB-NB model (top). The bottom plot magnifies the top plot for data values greater than* 50*. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

6.4.2 Aquatic science

The results for Aquatic Science citation data when fitted with the negative binomial, SVA and SVB models are given in Table 6.7.

Table 6.7: *Results for Aquatic Science.*

Models	Number of observations outside the boundaries	Threshold	Number of observations beyond the threshold
Negative binomial	13	$c = 167$	3
SVA Poisson-NB	27	$c = 168$	3
SVA NB-Poisson	12	$c = 171$	3
SVA NB-NB	14	$c = 202$	2
SVB Poisson-NB	12	$c = 178$	2
SVB NB-NB	13	$c = 169$	3

The Christmas tree plot for the negative binomial model indicates that this model is suitable for Aquatic Science (see Figure 6.23).

When fitted with the SVA Poisson-NB model, the results show that 27 points are outside the boundaries and the Christmas tree plot given in Figure 6.24 indicates that this model is unsuitable for Aquatic Science.

The Christmas tree plots for the SVA NB-Poisson and SVA NB-NB in Figures 6.25 and 6.26 show that these models are suitable as approximately 90% of the observed median adjusted counts are within the fluctuation boundaries. This is also observed for the SVB Poisson-NB and SVB NB-NB models (see Figures 6.27 and 6.28).

The results obtained for Aquatic Science are consistent with those obtained using randomised quantile residual plots in Section 5.4.2, except for the SVA Poisson-NB model, as the Christmas tree plot for SVA Poisson-NB model clearly shows larger variation in the data than expected from the null model.

The randomised quantile residual plot for the SVA Poisson-NB model gives evidence of large positive residuals, which is consistent with the Christmas tree plot as there are large observed counts beyond the threshold. If we enlarge the randomised quantile residual plot for SVA Poisson-NB (see Figure 6.22), it is clear that the points vary along the line. For instance, there is a slight dip when $-2 \leq z \leq 0$ and slight ascent thereafter. This dip corresponds to the cumulative distribution of the first 4 points, as we observed $P(c \leq 5) = 0.408$ and the corresponding quantile, $z = -0.23$ but under the SVA Poisson-NB model, $P(c \leq 5) = 0.44$ and the corresponding quantile, $z = -0.15$. The slight ascent at about $0 \leq z \leq 1$ corresponds to the 50^{th} to 84^{th} percentile, as more counts are observed than expected under the null model when $7 \leq c \leq 19$. Whilst the

differences for individual counts are apparent from a Christmas tree plot, this is not the case in a randomised quantile residual plot as the cumulative distribution functions of the observed counts are close to that under the fitted SVA Poisson-NB model.

Hence, unlike the Christmas tree plots, the randomised quantile residual plots are insensitive to individual data values, and the points will be close to the line if the cumulative distribution is on target.

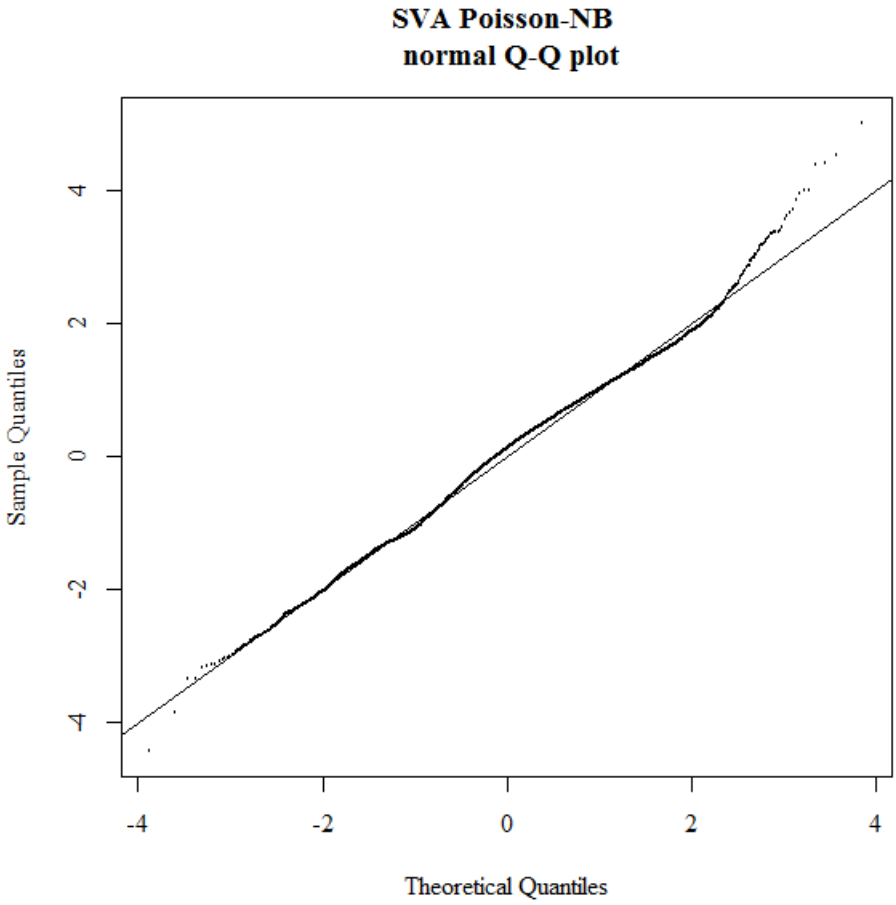

Figure 6.22: *An enlarged randomised quantile residual plot for Aquatic Science when fitted with the SVA Poisson-NB model.*

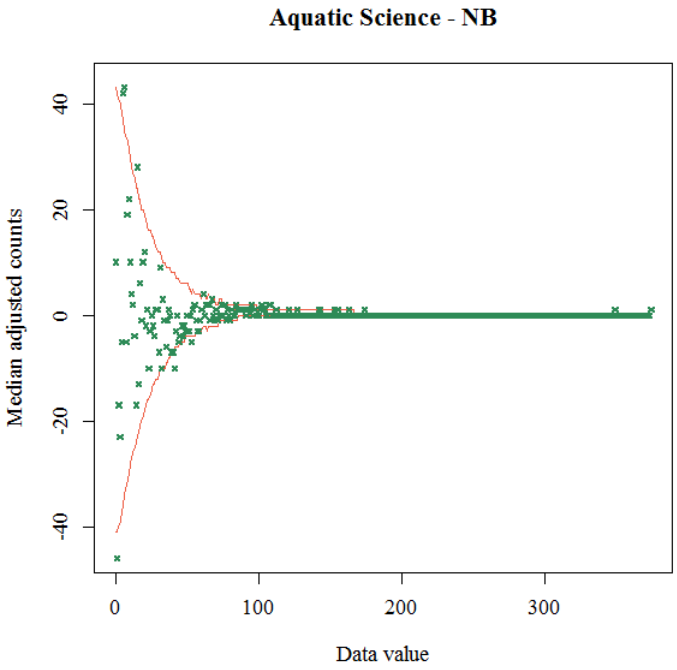

Figure 6.23: *A Christmas tree plot for Aquatic Science when fitted with the negative binomial model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

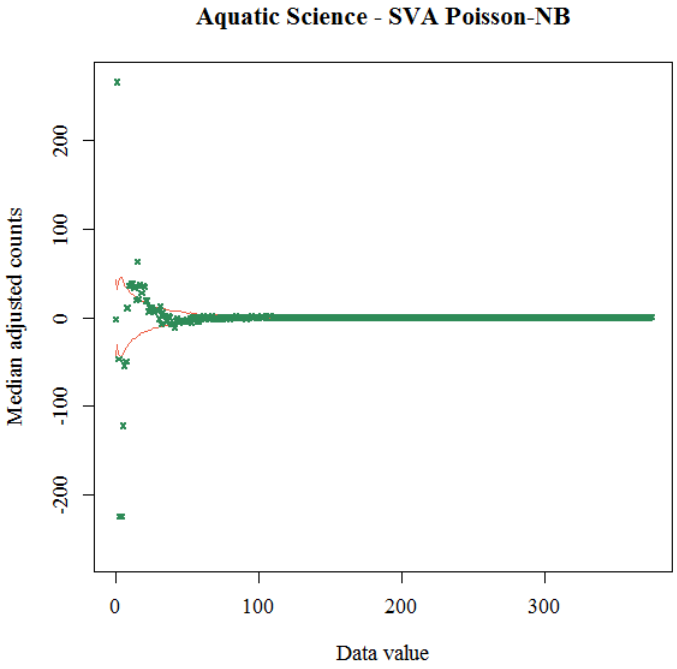

Figure 6.24: *A Christmas tree plot for Aquatic Science when fitted with the SVA Poisson-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

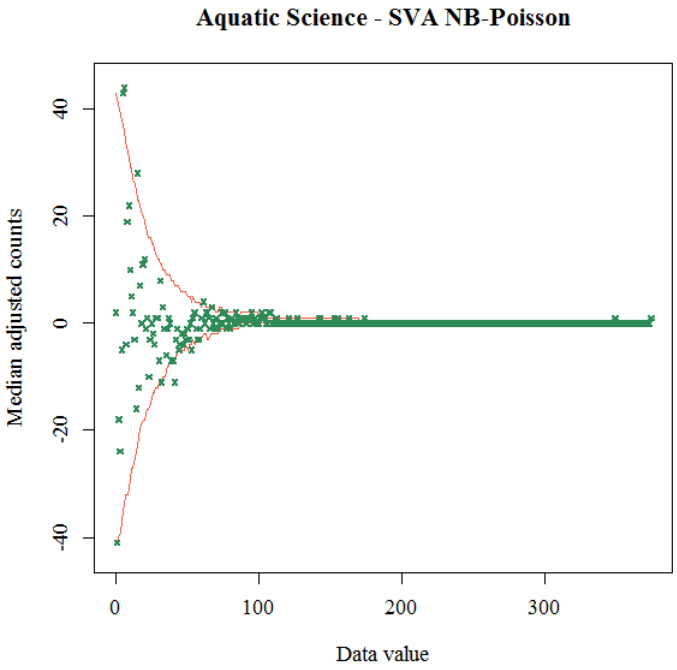

Figure 6.25: *A Christmas tree plot for Aquatic Science when fitted with the SVA NB-Poisson model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

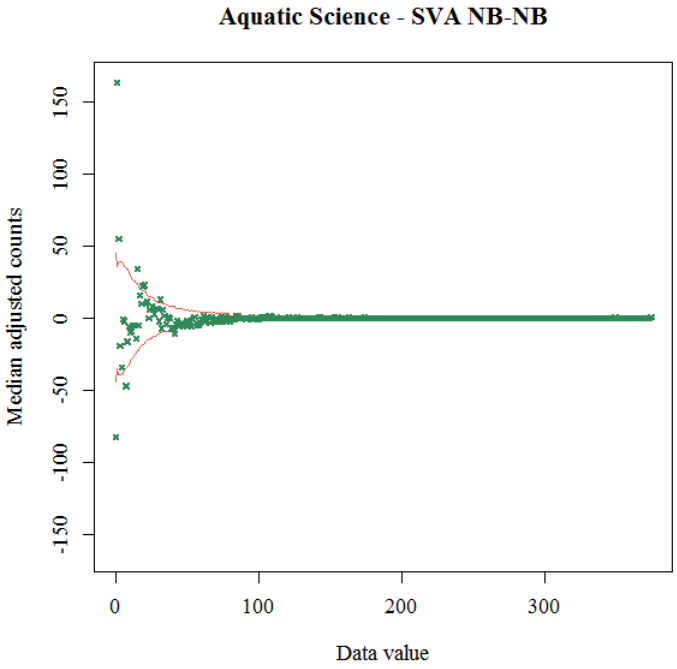

Figure 6.26: *A Christmas tree plot for Aquatic Science when fitted with the SVA NB-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

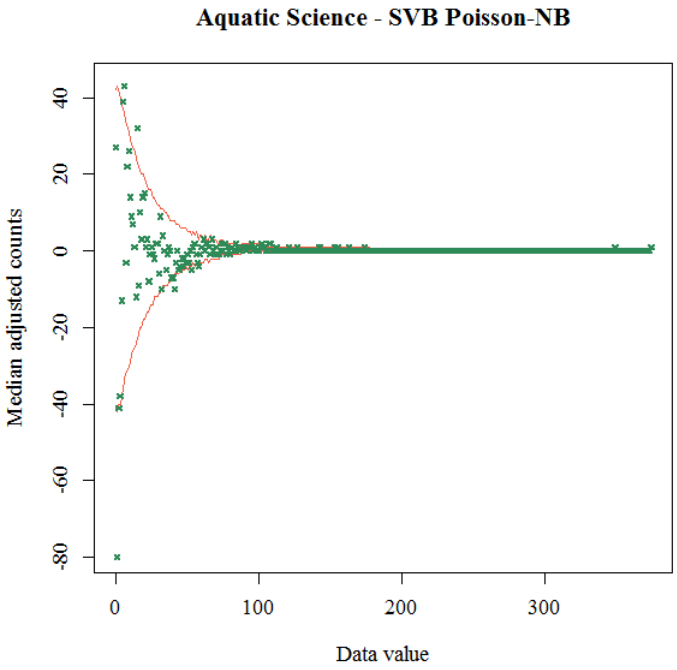

Figure 6.27: *A Christmas tree plot for Aquatic Science when fitted with the SVB Poisson-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

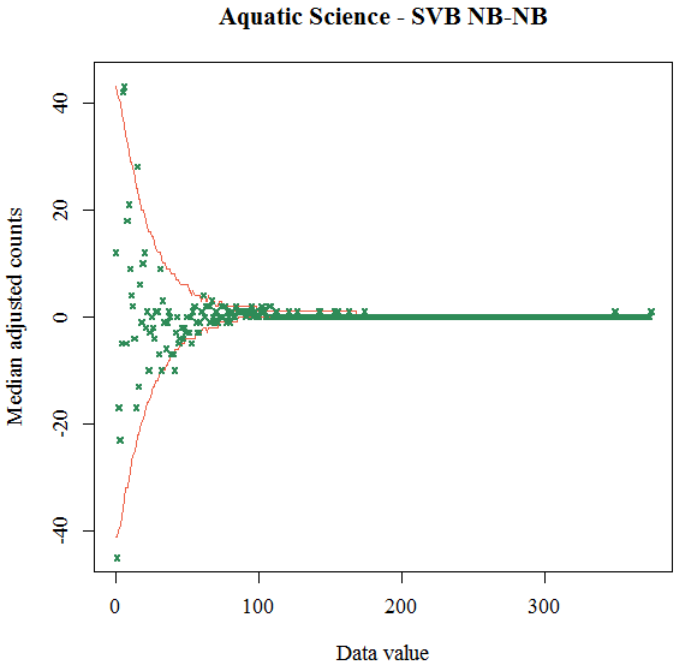

Figure 6.28: *A Christmas tree plot for Aquatic Science when fitted with the SVB NB-NB model. The orange lines denote the* 90% *fluctuation intervals while the green crosses are the observed median adjusted counts.*

6.5 Summary

Christmas tree plots are used to indicate how well the full range of count data scatter relative to a fitted distribution. These plots give evidence of whether the observed counts are consistent with those expected from the null model (Wilson and Einbeck, 2015). These plots illustrate an alternative model validation technique to AIC or BIC. Applying them to the citation models used in this thesis also enables comparison with the conclusions obtained from the randomised quantile residuals plots in the previous chapter. The Christmas tree plots for Tourism show that all models fitted are suitable, while the plots for Applied Mathematics indicate that they are unsuitable. For Aquatic Science, all models apart from the SVA Poisson-NB models are suitable. The Christmas tree plots clearly illustrates the observed citation counts, in particular they show the presence of large citation counts, which are untypical and beyond the threshold value of the fitted models in the subject areas investigated.

Randomised quantile residual plots versus Christmas tree plots

In this thesis, two diagnostic plots, randomised quantile residual plots and Christmas tree plots, are used to assess model fits. Recall that the quantile residuals are obtained by mapping the inverse of the fitted distribution function at each observed value to the equivalent standard normal deviate (Dunn and Smyth, 1996), while the fluctuation intervals in the test proposed by Wilson and Einbeck (2015) are computed here via the 5^{th} and 95^{th} quantiles of the Poisson-Binomial distribution for each response value.

Both diagnostic plots are consistent in terms of the underestimation of larger values by some fitted models. However, unlike the Christmas tree plots, the randomised quantile residual plots are insensitive to the under or overestimation of individual counts under the null model. This is because if the cumulative distribution of the observed counts are approximately equal to those under the null model, then the points will lie approximately along the reference line. In addition, it is difficult to read the randomised quantile residual plots when the difference in quantiles are very small or when the data set has very large sample size. It has been recommended that four realisations should be presented for the randomised quantile residual plots, where inconsistent patterns are ignored thereafter (Dunn and Smyth, 1996), but this is not necessary for the Christmas tree plots as randomisation is not used. Both diagnostic plots are suitable for assessing model fits, but have very different methods of interpretation. Whilst the randomised quantile residual plots may be interpreted using the analogies with Q-Q plots, this may still be tricky for a novice, and hence it may be more

straightforward to interpret the Christmas tree plots.

Chapter 7

Conclusions

This chapter provides a summary of the key findings and details the novel contributions of this thesis. The limitations of the research and suggestions for future studies are also discussed.

7.1 Key findings

This thesis introduces two variants of compound models, SVA and SVB. The variant models considered are generated from Poisson and negative binomial distributions. This thesis focuses on the development of these models. Detailed descriptions of their properties and the method of computation of their probabilities using the R software are given in Chapter 3. In Chapter 4, we used simulation studies to compare simulated data from SVA, SVB and standard compound distributions. We found that the generating model is not always selected by AIC or BIC. For example, for simulated data from SVA distributions, the AIC/BIC criteria may select the NB hurdle/ZINB models. Moreover, for some fixed parameters, SVA distributions are similar to standard compound distributions.

In Chapter 5, we showed that citation counts may be viewed as two generations that are not completely independent. This justifies the application of the SVA and SVB models in citation analysis. Moreover, both variant models have practical interpretations in citation analysis. The SVB models may be associated with the unusual 'sleeping beauties' in citation terms. The SVA models may be more useful as they replicate the well known "rich get richer" effect in scientometrics to some extent. The variant models were applied to citation data across various fields. Analyses were carried out using two sets of citation data. The first is covariate free while the second has two covariates: the number of authors and the number of affiliated countries.

In this thesis, model fits were assessed using the standard AIC and BIC criteria. Comparing the AIC and BIC criteria, since citation data often involve

large counts, the BIC should be used to avoid over fitting. We fitted standard compound models such as the Neyman type A model to citation data, but the associated AIC/BIC are much larger than those for SVA/SVB models, indicating that compound models are inappropriate for modelling citation data.

Apart from AIC/BIC, two diagrammatic methods were used to assess the appropriateness of models. The first uses randomised quantile residuals, in which the fitted distribution function is inverted at each response value to find its equivalent standard normal quantile. This is then presented as a normal probability plot of quantile residuals. However, the randomised quantile residuals are calculated using cumulative distribution functions, and hence are insensitive to individual values as the plotted points will still be very close to the reference line if the cumulative distribution is consistent with that under the fitted model. The second test uses fluctuation intervals to check for number inflation or deflation relative to a count model, with a Christmas tree plot (see Chapter 6). Despite using very different approaches, both methods lead to useful diagnostic devices to check the adequacies of models. Although Christmas tree plots have a more straightforward interpretation, especially for a statistical novice, they are more time consuming compared to the randomised quantile residual plots in the presence of covariates and large data values. However, unlike the randomised quantile residual plots, the Christmas tree plots are able to show if each individual count is consistent with that under the fitted model.

Referring back to the research questions listed in Section 1.2, the answers to these are mainly presented in Chapters 5 and 6. The answers to the research questions can be summarised as follows:

(i) **Are compound models appropriate for modelling citation data?**
In Section 5.3, the fits of various models, including two compound models, Neyman type A and Polya Aeppli are assessed for covariate free citation counts. In no case were these compound models selected by AIC/BIC, as the discretised lognormal and variants of compound models are preferred. Moreover the AIC/BIC of compound models are generally much larger than the other tested models, indicating that the compound models may be inappropriate for modelling covariate free citation data.

(ii) **Are the proposed variants of compound models suitable for citation analysis?**
This was investigated in Section 5.2. Citation counts of sets of articles were collected in two consecutive time periods, so that the counts in these two periods may be viewed as two generations. We found evidence that these two generations are not completely independent, indicating that the

proposed variants of compound models are suitable.

Further to the answers in (i), in Section 5.3, we found that amongst the models analysed (including negative binomial, Neyman type A, Polya Aeppli, SVA and SVB models), the variant models, especially SVB NB-NB and SVA NB-NB returned the lowest AIC/BIC for some subject areas, indicating that they are suitable. In addition, the applications of the variant models also provide practical interpretations in citation terms, thus are suitable for modelling citation data.

(iii) **Are randomised quantile residual plots and Christmas tree plots more useful than AIC and BIC for model validation in citation analysis?**
Both diagrammatic methods are useful model validation techniques. In particular, the presence of citation counts that are larger than expected under the fitted model are detected in both plots. However, this is not the case for smaller citation counts. Although the Christmas tree plots are able to reveal cases where there are more or less counts than expected under the fitted model for some smaller data values (see Section 6.4), randomised quantile residual plots cannot, especially when the observed cumulative distributions are close to those under the fitted model.

(iv) **Can the proposed variants of compound models be extended to incorporate covariates?**
Examples of the code for fitting covariate free SVA or SVB models are presented in Section 3.5. This may be extended by adding extra parameters to incorporate covariates and an example is presented in Section 5.4, where two covariates are added to the model for analysis of citation data sets.

7.2 Novel contributions

The first main contribution of this thesis is the introduction and development of variants of compound models, with detailed descriptions of their properties.

This thesis also provides novel contributions to the scientometrics community, by assessing the new SVA and SVB models (see Chapter 3 and 5), for citation analyses. We showed that the variant models, especially SVA, are suitable to model citation data, and their interpretation gives new insights into the citation process.

We performed model assessment using randomised quantile residual plots and Christmas tree plots. The randomised quantile residuals have been applied to

generalised linear models (Smyth et al., 2015), but are further extended to variants of compound models in this thesis. In addition, we extended the application of the test for number inflation and deflation to a much larger range, from zero to hundreds. Previously, this test and its associated Christmas tree plots were used when the relative fitted distribution is a Poisson or zero-modified Poisson model (Wilson and Einbeck, 2015). Here, the test is extended to variants of compound models.

7.3 Limitations and further work

In this thesis, the SVA and SVB distributions considered consist of two generations, where one is Poisson and the other is negative binomial, or both generations are negative binomial. It may be extended by considering other count distributions, so that it may be applied to either under or over-dispersed data. The model fitting algorithm is currently based on maximum likelihood estimation method via the *optim* command, but in some cases the model fails to converge. Currently, the fitting of these models is only for covariate free data or data with two covariates. Thus it is necessary to adjust the model fitting process manually for other tasks. It may be beneficial to automate this process and to develop an R package to ease future analysis.

Here, the citation analyses incorporates only two covariates (number of authors and number of affiliated countries), but this could be extended to investigate other factors affecting citation counts. Given the variability of citing practices across different fields, future research could include the incorporation of an offset term into the models to account for the differences in the mean number of citations across fields. The data sets used in this thesis are mainly citation counts of journal articles, but it may be advantageous to apply the variant models to other data types, such as citation counts of textbooks or readership counts from reference managers like Mendeley.

Some of the Christmas tree plots showed the presence of citation counts that are larger than expected under the null model. These are commonly far beyond the point where the fluctuation intervals merge. These are extreme points for the hypothesised model. Thus, it may be useful to incorporate extreme value theory for citation data in future research or to otherwise seek to explain these values.

In this thesis, the variant models are mostly applied to citation data and briefly applied to biodosimetry data. Although there is evidence that these models are suitable for the former, they are not suitable for the latter. This is unsurprising as applications of the variant models to biodosimetry data lack practical interpretation. It may be beneficial to extend the application of these models to other

sources of count data, for which the SVA or SVB models may be suitable.

References

Adler, A. (2014). *Delaporte: Statistical Functions for the Delaporte Distribution*. R package version 2.2-2. URL http://CRAN.R-project.org/package=Delaporte.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest.

Albarrán, P. and Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, 62(1):40–49.

Baglivo, J. A. (2005). *Mathematica laboratories for mathematical statistics: Emphasizing simulation and computer intensive methods*. SIAM, Philadelphia, ASA, Alexandria, VA.

Becker, M. P., Yang, I., and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research*, 6(1):38–54.

Bookstein, A. (1990). Informetric distributions, part II: Resilience to ambiguity. *Journal of the American Society for Information Science*, 41(5):376–386.

Bookstein, A. (2001). Implications of ambiguity for scientometric measurement. *Journal of the American Society for Information Science and Technology*, 52(1):74–79.

Bornmann, L. and Daniel, H. D. (2006). Selecting scientific excellence through committee peer review - A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68(3):427–440.

Bornmann, L. and Daniel, H. D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by angewandte chemie international edition, or rejected but published

elsewhere. *Journal of the American Society for Information Science and Technology*, 59(11):1841–1852.

Bornmann, L. and Daniel, H. D. (2016). Count regression models in informetrics. *Journal of Informetrics*, 10(1):29–30.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, 103(1):213–228.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag New York, New York, 2nd edition.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

Cameron, A. C. and Trivedi, P. K. (1999). Essentials of count data regression. In *A Companion to Theoretical Econometrics*, pages 331–348. Blackwell Publishing Ltd, Malden, MA, USA.

Case, D. O. and Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7):635–645.

Ceppellini, B. R., Siniscalco, M., and Smith, C. A. B. (1955). The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics*, 20(2):97–115.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-Binomial and Conditional Bernoulli distributions. *Statistica Sinica*, 7(4):875–892.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, 70(1):269–274.

Daley, D. and Vere-Jones, D. (2003). *An introduction to the theory of point processes*. Probability and its Applications. Springer-Verlag, New York.

Dean, C., Lawless, J. F., and Willmot, G. E. (1989). A mixed poisson-inverse-gaussian regression model. *Canadian Journal of Statistics*, 17(2):171–181.

DeLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics*, pages 599–609. Springer New York, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38.

Di Giorgio, M., Edwards, A. A., Moquet, J. E., Finnon, P., Hone, P. A., Lloyd, D. C., Kreiner, A. J., Schuff, J. A., Tajal, M. R., Vallerga, M. B., López, F. O., Burlón, A., Debray, M. E., and Valda, A. (2004). Chromosome aberrations induced in human lymphocytes by heavy charged particles in track segment mode. *Radiation Protection Dosimetry*, 108(1):47–53.

Didegah, F. and Thelwall, M. (2013a). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, 64(5):1055–1064.

Didegah, F. and Thelwall, M. (2013b). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, 7(4):861–873.

Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8):897–899.

Dobbie, M. and Welsh, A. (2001). Models for zero-inflated count data using the Neyman type A distribution. *Statistical Modelling*, 1(1):65–80.

Dodge, Y. (2008). Coefficient of kurtosis. In *The Concise Encyclopedia of Statistics*, pages 91–92. Springer New York, New York, NY.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Dziak, J. J., Coffman, D. L., Lanza, S. T., and Runze, L. (2012). Sensitivity and specificity of information criteria. Technical Report 119, The Pennsylvania State University.

Einbeck, J. and Wilson, P. (2016). A diagnostic plot for assessing model fit in count data models. In Dupuy, J.-F. and Josse, J., editors, *Proceedings of the 31st International Workshop on Statistical Modelling*, volume 1, pages 103–108, Rennes.

Eom, Y.-H. and Fortunato, S. (2011). Characterizing and Modeling Citation Dynamics. *PLoS ONE*, 6(9):e24926.

Evans, T. S., Hopkins, N., and Kaube, B. S. (2012). Universality of performance indicators based on citation and reference counts. *Scientometrics*, 93(2):473–495.

Feather, J. and Sturges, P. (2003). *International Encyclopedia of Library and Information Science*. Routledge, Taylor & Francis group, London and New York, 2nd edition.

Foster, S. D. and Bravington, M. V. (2013). A PoissonGamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, 20(4):533–552.

Garcia, J. M. G. (2011). A fixed-point algorithm to estimate the YuleSimon distribution parameter. *Applied Mathematics and Computation*, 217(21):8560–8566.

Glanzel, W. (2007). Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1(1):92–102.

Goffman, W. and Newill, V. A. (1964). Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204(4955):225–228.

Grimmett, G. and Welsh, D. (1986). *Probability: An introduction*. Oxford University Press, New York, U.S.A.

Hankin, R. K. S. (2006). Additive Integer Partitions in R. *Journal of Statistical Software*, 16(1):1–3.

Harzing, A.-w. and Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804.

HEFCE (2009). Research Excellence Framework: Second consultation on the assessment and funding of research. Technical report.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press, 2nd, revis edition.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Hogg, R. V. and Craig, A. T. (1995). *Introduction to Mathematical Statistics*. Prentice-Hall, Inc, New Jersey, 5th edition.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis*, 59(1):41–51.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Hurvich, C. M. and Tsai, C. L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, 51(3):1077–1084.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distribution*. Wiley-Interscience, 3rd edition.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous univariate distributions*. Wiley-Interscience, 2nd edition.

Kadane, J. B. and Lazar, N. a. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kinney, A. L. (2007). National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Sciences*, 104(46):17943–17947.

Kinney, J. J. (1997). *Probability: An introduction with statistical applications*. John Wiley & Sons, Inc, Canada.

Kretschmer, H. and Rousseau, R. (2001). Author inflation leads to a breakdown of Lotka's law. *Journal of the American Society for Information Science and Technology*, 52:610–614.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lee, Y. G., Lee, J. D., Song, Y. I., and Lee, S. J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics*, 70(1):27–39.

Lehmann, S., Lautrup, B., and Jackson, A. D. (2003). Citation networks in high energy physics. *Physical Review E*, 68(2):026113.

Leydesdorff, L. (2013). An evaluation of impacts in "Nanoscience & nanotechnology": Steps towards standards for citation analysis. *Scientometrics*, 94(1):35–55.

Leydesdorff, L. and Milojević, S. (2012). Scientometrics. In *The International Encyclopedia of Social and Behavioral Sciences*, chapter Science an, pages 1–20. Elsevier Ltd.

Liu, N. C. (2009). The story of academic ranking of world universities. *International Higher Education*, 54(4):2–3.

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268):847.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323.

Lüders, R. (1934). Die statistik der seltenen ereignisse. *Biometrika*, 26(1-2):108–128.

Ma, Y., Genton, M. G., and Parzen, E. (2011). Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, 63(2):227–243.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science (New York, N.Y.)*, 159(3810):56–63.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Mullahy, J. (1997). Heterogeneity , excess zeros , and the structure of count data models. *Journal of Applied Econometrics*, 12(3):337–350.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, 10(1):35–57.

Nomaler, Ö., Frenken, K., and Heimeriks, G. (2013). Do more distant collaborations have more citation impact? *Journal of Informetrics*, 7(4):966–971.

O'Keeffe, A. G., Tom, B. D. M., and Farewell, V. T. (2013). Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Statistics in medicine*, 32(4):600–619.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P., and Rothkamm, K. (2016). Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal*, 58(2):259–279.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5207–5211.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics*, 4(3):358–364.

Price, D. J. D. S. (1951). Quantitative measures of the development of science. *Archives Internationales dHistoire des Sciences*, 4(14):85–93.

Price, D. J. D. S. (1965). Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Puig, P. and Barquinero, J. F. (2011). An application of compound Poisson modelling to biological dosimetry. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2127):897–910.

Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101(473):332–340.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Radicchi, F. and Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, 97(3):627–637.

Radicchi, F., Fortunato, S., and Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272.

Rauhvargers, A. (2011). Global university rankings and their impact. Technical report, European University Association, Brussels, Belgium.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134.

Ridout, M., Demetrio, C. G., and Hinde, J. (1998). Models for count data with many zeros. In *International Biometric Conference*, number 19, pages 179–192.

Rigby, R. A., Stasinopoulos, D. M., and Akantziliotou, C. (2008). A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics and Data Analysis*, 53(2):381–393.

Rigby, R. A., Stasinopoulos, D. M., and Lane, P. W. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(3):507–554.

Romm, H., Ainsbury, E., Barnard, S., Barrios, L., Barquinero, J., Beinke, C., Deperas, M., Gregoire, E., Koivistoinen, A., Lindholm, C., Moquet, J., Oestreicher, U., Puig, R., Rothkamm, K., Sommer, S., Thierens, H., Vandersickel, V., Vral, A., and Wojcik, A. (2013). Automatic scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 756(1-2):174–183.

Rousseau, R. (1992). Breakdown of the robustness property of Lotka's Law: The case of adjusted counts for multiauthorship attribution. *Journal of the American Society for Information Science*, 43(10):645–647.

Sangwal, K. (2013). Comparison of different mathematical functions for the analysis of citation distribution of papers of individual authors. *Journal of Informetrics*, 7(1):36–49.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society. Series A (General)*, 137(1):25–34.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42(3-4):425–440.

Simon, S. L., Bouville, A., and Kleinerman, R. (2010). Current use and future needs of biodosimetry in studies of long-term health risk following radiation exposure. *Health Physics*, 98(2):109–117.

Skellam, J. G. (1952). Studies in statistical ecology. *Biometrika*, 39(3-4):346–362.

Smith, S., Ward, V., and House, A. (2011). 'Impact' in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. *Research Policy*, 40(10):1369–1379.

Smyth, G., Hu, Y., Dunn, P., Phipson, B., and Chen, Y. (2015). *statmod: Statistical Modeling*. R package version 1.4.21. URL http://CRAN.R-project.org/package=statmod.

Sud, P. and Thelwall, M. (2016). Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*, 67(8):1849–1857.

Thelwall, M. (2016a). Are there too many uncited articles? Zero inflated variants of the discretised lognormal and hooked power law distributions. *Journal of Informetrics*, 10(2):622–633.

Thelwall, M. (2016b). Citation count distributions for large monodisciplinary journals. *Journal of Informetrics*, 10(3):863–874.

Thelwall, M. (2016c). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, 10(2):336–346.

Thelwall, M. and Wilson, P. (2014a). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, 8(4):824–839.

Thelwall, M. and Wilson, P. (2014b). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, 8(4):963–971.

Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, 59(3):467–472.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer, New York, 4th edition.

Vieira, E. S. and Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of science for a typical university. *Scientometrics*, 81(2):587–600.

Vieira, E. S. and Gomes, J. a. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, 4(1):1–13.

Virsik, R. P. and Harder, D. (1981). Statistical interpretation of the overdispersed distribution of radiation-induced dicentric chromosome aberrations at high LET. *Radiation Research*, 85(1):13.

von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. Teubner, Leipzig.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2):228–243.

Willmot, G. E. (1987). The Poisson-Inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 1987(3-4):113–127.

Wilson, P. (2008). *Variations of Cox-type tests and their application to models for count data with modified zeros*. PhD thesis, National University of Ireland Galway.

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number-inflation or number-deflation. In Friedl, H. and Wagner, H., editors, *Proceedings of the 30th International Workshop on Statistical Modelling*, volume 2, pages 299–302, Linz, Austria.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero-modification in count data regression. In Dupuy, J.-F. and Josse, J., editors, *Proceedings of the 31st International Workshop on Statistical Modelling*, volume 1, pages 325–330, Rennes, France.

Yule, U. (1925). A mathematical theory of evolution, based on the conclusions of II. - a mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Source Philosophical Transactions of the Royal Society of London. Series B*, 213:21–87.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8):1–25.

Zhang, H., Liu, Y., and Li, B. (2014). Notes on discrete compound Poisson model with applications to risk theory. *Insurance: Mathematics and Economics*, 59:325–336.

Zuur, A. F., Hilbe, J. M., and Ieno, E. N. (2013). *A beginner's guide to GLM and GLMM with R.* Highland Statistics Ltd., Newburgh, United Kingdom.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed effects models and extensions in ecology with R.* Springer-Verlag New York, New York; London.

Appendix A

List of publications

1. Low, W. J., Wilson, P. and Thelwall, M. (2016). Stopped sum models and proposed variants for citation data. *Scientometrics*, 107(2):369-384.

2. Low, W. J., Wilson, P. and Thelwall, M. (2016). Variations of compound models. In Dupuy, J.-F. and Josse, J., editors, *Proceedings of the 31st International Workshop on Statistical Modelling*, volume 2, pages 67-70, Rennes, France.

3. Low, W. J., Wilson, P. and Thelwall, M. (2015). Stopped sum models for citation data, In Salah, A. A., Tonta, Y., Salah, A. A. A., Sugimoto, C., and Al., U., editors, *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*, pages 184-195, Boazii University, Istanbul, Turkey.

4. Low, W. J., Wilson, P. and Thelwall, M. (2015). Stopped sum model variants. Poster presentation at *the 8th International Conference of the ERCIM WG on Computational and Methodological Statistics*, University of London, U.K.

Appendix B

Expectations and variances of compound models

Since compound distributions can be viewed as the sum of a random number of independent random variables, if we denote S_N as a compound A-B distribution, where $S_N = X_1 + X_2 + \cdots + X_N$, where $N, X_1, X_2, \cdots$ are independent random variables, where N is distributed by distribution A and $X_1, X_2, \cdots$ are identically distributed with distribution B, then the expectation of S is:

$$\begin{aligned} E(S) &= E(E(S|N)) \qquad \text{by Adam's law} && \text{(B.1)} \\ &= E(E(X_1 + X_2 + X_3 + \cdots + X_N|N)) && \text{(B.2)} \\ &= E(NE(X)) && \text{(B.3)} \\ &= E(N) \cdot E(X) && \text{(B.4)} \end{aligned}$$

Therefore expectations of the compound distributions considered are:

Table B.1: *Expectations of the compound distributions considered.*

Distributions	E(X)
Neyman type A	$\lambda\phi$
Compound Poisson-NB	$\lambda\mu$
Compound NB-Poisson	$\lambda\mu$
Compound NB-NB	$\mu_1\mu_2$

The variance of S is:

$$Var(S) = E(N)Var(X) + Var(N)\left(E(X)\right)^2 \qquad \text{(B.5)}$$

Hence the variances of the compound distributions considered are:

Table B.2: *Variances of compound distributions considered.*

Distributions	Var(X)
Neyman type A	$\lambda\phi(1+\lambda)$
Compound Poisson-NB	$\mu\lambda\left(1+\mu+\frac{\mu}{\alpha}\right)$
Compound NB-Poisson	$\mu\lambda\left(1+\lambda+\frac{\mu\lambda}{\alpha}\right)$
Compound NB-NB	$\mu_1\mu_2\left(1+\mu_2+\frac{\mu_2}{\alpha_2}+\frac{\mu_1\mu_2}{\alpha_1}\right)$

As mentioned in Sections 3.4.3 and 3.4.4, the expectations and variances can also be derived using the mgf or pgf and similar results to Tables B.1 and B.2 will be obtained. Note that if a distribution X has pgf, $G_X(t)$, and mgf, $M_X(t)$, then

$$G_X(e^t) = M_X(t) \tag{B.6}$$

Appendix C

Results for citation analysis with no covariates

Table C.1: *Results obtained when fitted with negative binomial model.*

Subject	Coefficients		Standard error	
	μ	θ	μ	θ
Visual	0.660	0.172	0.028	0.008
Tourism	21.535	0.978	0.903	0.054
Soil	16.930	0.743	0.304	0.016
Marketing	26.129	0.633	0.844	0.021
Literature	0.792	0.315	0.024	0.013
Horticulture	16.718	0.830	0.343	0.021
History	2.900	0.300	0.079	0.008
Genetics	39.234	0.610	0.716	0.011
Ecology	25.022	0.864	0.387	0.017
Developmental	35.448	0.930	0.553	0.018
Biochem	28.808	0.837	0.452	0.016
Accounting	25.894	0.644	0.952	0.025
AppliedMaths	11.715	0.499	0.239	0.010
Urology	19.391	0.513	0.388	0.010
StatsProb	16.931	0.538	0.332	0.010
Rehab	9.285	0.231	0.277	0.005
Oncology	40.234	0.547	0.803	0.011
Logic	13.404	0.526	0.280	0.011
Dermatology	8.074	0.646	0.185	0.017
Algebra	5.746	0.904	0.283	0.065

Table C.2: *Results obtained when fitted with SVA Poisson-NB model.*

Subject	Coefficients			Standard error		
	λ	μ	α	λ	μ	α
Visual	0.275	1.605	0.336	0.009	0.098	0.027
Tourism	3.222	18.772	0.567	0.172	1.051	0.040
Soil	2.268	16.087	0.563	0.042	0.349	0.015
Marketing	2.627	24.968	0.430	0.082	1.009	0.018
Literature	0.401	1.185	0.332	0.010	0.059	0.024
Horticulture	2.522	15.146	0.539	0.057	0.399	0.017
History	0.748	4.081	0.273	0.015	0.158	0.011
Genetics	2.707	38.784	0.498	0.049	0.807	0.010
Ecology	2.520	24.168	0.792	0.046	0.407	0.018
Developmental	4.027	31.856	0.600	0.092	0.626	0.014
Biochem	3.210	26.598	0.613	0.063	0.499	0.014
Accounting	2.459	25.358	0.497	0.088	1.104	0.023
AppliedMaths	1.680	12.203	0.392	0.028	0.310	0.010
Urology	1.797	20.692	0.499	0.030	0.455	0.012
StatsProb	2.125	16.618	0.357	0.035	0.422	0.009
Rehab	0.828	14.562	0.368	0.016	0.451	0.011
Oncology	2.337	41.675	0.535	0.043	0.883	0.011
Logic	1.669	14.207	0.488	0.030	0.339	0.013
Dermatology	1.794	7.442	0.369	0.037	0.244	0.014
Algebra	1.904	4.458	0.368	0.096	0.368	0.041

Table C.3: *Results obtained when fitted with SVA NB-Poisson model.*

Subject	Coefficients			Standard error		
	μ	α	λ	μ	α	λ
Visual	0.661	0.172	0.000	-	-	-
Tourism	21.525	0.977	0.008	0.930	0.061	0.228
Soil	16.874	0.737	0.065	0.310	0.017	0.064
Marketing	26.019	0.625	0.121	0.852	0.022	0.110
Literature	0.792	0.315	0.000	-	-	-
Horticulture	16.709	0.829	0.009	0.349	0.023	0.072
History	2.900	0.300	0.000	-	-	-
Genetics	38.965	0.598	0.279	0.722	0.011	0.080
Ecology	24.729	0.842	0.314	0.399	0.018	0.101
Developmental	34.564	0.864	0.900	0.574	0.019	0.132
Biochem	28.084	0.786	0.748	0.465	0.016	0.106
Accounting	25.661	0.628	0.259	0.964	0.026	0.141
AppliedMaths	11.715	0.500	0.000	-	-	-
Urology	19.469	0.510	0.000	-	-	-
StatsProb	16.933	0.538	0.000	-	-	-
Rehab	9.282	0.231	0.000	-	-	-
Oncology	39.938	0.536	0.328	0.810	0.011	0.089
Logic	13.366	0.526	0.000	-	-	-
Dermatology	8.063	0.645	0.012	0.189	0.018	0.045
Algebra	5.736	0.902	0.012	0.298	0.068	0.109

Table C.4: *Results obtained when citation data are fitted with SVA NB-NB model.*

Subject	Coefficients				Standard error			
	μ_1	α_1	μ_2	α_2	μ_1	α_1	μ_2	α_2
Visual	0.601	0.194	0.256	0.002	0.025	0.010	0.188	0.001
Tourism	13.482	1.299	8.255	0.103	2.052	0.139	2.300	0.046
Soil	13.777	0.825	3.460	0.039	0.712	0.024	0.788	0.014
Marketing	20.339	0.761	6.157	0.014	1.229	0.033	1.775	0.007
Literature	0.409	9.217	1.158	0.314	-	-	-	-
Horticulture	14.267	0.937	2.615	0.020	0.625	0.032	0.694	0.009
History	1.257	0.748	3.119	0.118	0.116	0.123	0.279	0.023
Genetics	24.305	0.805	15.847	0.044	1.067	0.020	1.526	0.008
Ecology	22.606	0.762	2.596	0.315	0.471	0.018	0.319	0.043
Developmental	17.955	1.521	17.733	0.121	0.875	0.056	1.148	0.015
Biochem	22.858	1.119	6.089	0.009	0.483	0.026	0.991	0.002
Accounting	12.928	0.873	14.025	0.121	2.789	0.100	3.226	0.046
AppliedMaths	8.198	0.635	4.278	0.030	0.393	0.021	0.586	0.007
Urology	15.491	0.564	4.595	0.032	0.765	0.014	0.920	0.010
StatsProb	10.505	0.769	7.213	0.029	0.365	0.022	0.734	0.005
Rehab	0.831	89.545	14.556	0.367	0.016	60.776	0.451	0.011
Oncology	25.496	0.685	16.330	0.047	1.356	0.017	1.822	0.010
Logic	11.594	0.564	2.191	0.020	0.572	0.016	0.679	0.010
Dermatology	1.833	41.253	7.395	0.362	-	-	0.239	0.011
Algebra	1.945	42.306	4.411	0.358	0.160	129.408	0.398	0.051

Table C.5: *Results obtained when citation data are fitted with SVB Poisson-NB model.*

Subject	Coefficients			Standard error		
	λ	μ	α	λ	μ	α
Visual	0.037	0.623	0.138	0.012	0.031	0.012
Tourism	1.413	20.122	0.747	0.340	1.019	0.065
Soil	0.108	16.823	0.721	0.066	0.314	0.020
Marketing	1.020	25.109	0.504	0.140	0.918	0.023
Literature	11.819	11.995	0.000	-	-	-
Horticulture	0.499	16.239	0.728	0.102	0.369	0.027
History	0.197	2.703	0.212	0.020	0.088	0.009
Genetics	0.426	38.808	0.569	0.080	0.737	0.013
Ecology	0.000	23.597	0.906	-	-	-
Developmental	2.559	32.889	0.688	0.164	0.616	0.020
Biochem	0.688	28.120	0.761	0.128	0.479	0.020
Accounting	0.339	25.554	0.598	0.154	0.986	0.031
AppliedMaths	0.279	11.436	0.436	0.041	0.253	0.012
Urology	0.024	19.366	0.509	0.035	0.391	0.012
StatsProb	0.780	16.158	0.411	0.056	0.365	0.011
Rehab	0.092	9.193	0.205	0.014	0.291	0.006
Oncology	0.000	45.662	0.540	-	-	-
Logic	0.038	13.366	0.517	0.036	0.283	0.014
Dermatology	0.598	7.476	0.470	0.067	0.210	0.021
Algebra	0.837	4.909	0.546	0.204	0.364	0.082

Table C.6: *Results obtained when citation data are fitted with SVB NB-NB model.*

Subject	Coefficients				Standard error			
	μ_1	α_1	μ_2	α_2	μ_1	α_1	μ_2	α_2
Visual	0.598	0.191	0.063	0.001	-	-	-	-
Tourism	14.748	0.353	6.785	1.168	2.907	0.138	2.737	0.184
Soil	4.919	0.079	12.012	0.751	1.542	0.041	1.534	0.036
Marketing	8.353	0.035	17.776	0.755	2.121	0.018	1.867	0.032
Literature	4.652	2.707	3.851	0.000	-	-	-	-
Horticulture	3.816	0.045	12.902	0.914	1.062	0.023	1.041	0.030
History	1.078	0.379	1.823	0.072	0.130	0.021	0.160	0.013
Genetics	15.115	0.044	24.118	0.753	1.473	0.008	1.133	0.018
Ecology	3.355	0.022	21.666	0.926	0.741	0.009	0.742	0.021
Developmental	18.404	0.140	17.044	1.408	1.163	0.018	0.929	0.049
Biochem	5.789	0.008	23.019	1.106	0.998	0.002	0.466	0.025
Accounting	18.480	0.251	7.395	0.605	2.408	0.064	2.170	0.071
AppliedMaths	4.263	0.040	7.453	0.580	0.561	0.010	0.494	0.016
Urology	4.168	0.032	15.209	0.521	1.116	0.015	1.108	0.017
StatsProb	7.191	0.037	9.740	0.721	0.656	0.006	0.424	0.019
Rehab	5.710	0.000	25.741	0.197	-	-	-	-
Oncology	11.428	0.021	28.806	0.638	1.526	0.005	1.131	0.014
Logic	2.521	0.031	10.883	0.528	1.088	0.022	1.091	0.021
Dermatology	3.223	0.812	4.851	0.161	0.422	0.048	0.470	0.031
Algebra	2.481	1.255	3.265	0.227	1.055	0.283	1.093	0.144

Appendix D

Results for citation analysis with covariates

The estimated parameters and their associated standard errors of the models fitted in Section 5.4 are presented here.

Table D.1: *Results obtained when citation data are fitted with the standard negative binomial model. A dash '-' indicates that the model is unsuitable.*

Subject	Estimated coefficients				Standard errors			
	μ_{inter}	μ_{author}	$\mu_{countries}$	α	μ_{inter}	μ_{author}	$\mu_{countries}$	α
Applied maths	1.253	0.177	0.380	0.585	0.059	0.013	0.052	0.019
Aquatic	1.820	0.070	0.201	−0.087	0.027	0.005	0.019	0.016
Archeology	0.252	0.218	0.167	1.023	0.205	0.044	0.210	0.058
Biochemistry	1.649	0.089	0.316	0.383	0.057	0.007	0.035	0.025
Biomedical	1.843	0.100	0.312	0.818	0.072	0.010	0.053	0.021
Biophysics	2.363	0.045	0.124	0.090	0.034	0.005	0.023	0.015
Care planning	0.222	0.384	0.038	0.515	0.379	0.046	0.333	0.109
Neuroscience	2.539	0.035	0.162	−0.120	0.029	0.004	0.018	0.017
Chemical health	1.314	0.015	0.514	−0.106	0.232	0.030	0.178	0.103
Computer	1.464	0.076	0.441	0.696	0.093	0.017	0.080	0.025
Physics	1.768	0.097	0.326	0.597	0.056	0.009	0.045	0.018
Developmental	2.098	0.100	0.123	0.231	0.042	0.007	0.034	0.017
Earth	1.852	0.113	0.137	0.326	0.039	0.008	0.032	0.020
Education	1.310	0.177	0.217	0.721	0.093	0.013	0.088	0.019
Electronic	2.314	0.055	0.118	0.314	0.050	0.008	0.039	0.020
Environmental	2.510	0.062	0.121	0.089	0.033	0.006	0.024	0.015
Inorganic	2.206	0.039	0.154	0.207	0.039	0.006	0.030	0.015
Management	1.925	0.200	−0.057	0.691	0.135	0.034	0.126	0.034
Microbiology	2.254	0.050	0.168	0.022	0.029	0.004	0.019	0.016
Nuclear	1.879	0.018	0.109	0.767	0.047	0.005	0.041	0.021
Oral Surgery	1.865	0.018	0.266	−0.319	0.169	0.024	0.113	0.081
Pharmacology	1.562	0.076	0.288	0.435	0.062	0.008	0.044	0.025
Small Animals	0.656	0.232	0.202	0.656	0.138	0.023	0.102	0.054
Statistics	1.537	0.144	0.162	0.471	0.064	0.020	0.054	0.021

Table D.2: *Results obtained when citation data are fitted with the Neyman type A model. A dash '-' indicates that the model is unsuitable. Biochemistry and Condensed Matter Physics are excluded as this model is unsuitable for these subjects.*

Subject	Estimated coefficients				Standard error			
	μ_{inter}	μ_{author}	$\mu_{countries}$	d	μ_{inter}	μ_{author}	$\mu_{countries}$	d
Applied maths	14.635	−0.497	−0.991	3.603	-	-	-	-
Aquatic	3.603	0.827	3.189	5.355	0.260	0.042	0.193	0.038
Archeology	0.220	0.606	1.056	3.617	0.493	0.086	0.513	0.081
Biomedical	9.071	−0.235	5.839	3.167	-	-	-	-
Biophysics	11.104	0.451	1.729	5.470	-	-	-	-
Care planning	0.722	1.902	−0.758	4.559	1.287	0.165	1.272	0.188
Neuroscience	9.798	0.717	3.433	3.885	-	-	-	-
Chemical health	2.852	0.074	3.620	4.582	1.498	0.163	1.284	0.203
Computer	0.921	0.661	5.630	6.017	0.520	0.093	0.441	0.054
Developmental	5.372	1.982	1.540	6.554	0.429	0.066	0.408	0.044
Earth	3.201	1.442	2.913	6.160	0.276	0.062	0.273	0.048
Education	2.118	1.342	1.982	5.401	0.379	0.058	0.381	0.037
Electronic	10.457	0.401	2.233	6.074	-	-	-	-
Environmental	10.473	1.054	2.191	4.189	-	-	-	-
Inorganic	7.992	0.546	2.258	6.469	0.302	0.045	0.290	0.037
Management	6.066	1.860	0.114	6.618	0.847	0.190	0.814	0.078
Microbiology	7.326	0.811	3.236	7.201	0.381	0.047	0.257	0.047
Nuclear	11.144	−0.090	−0.419	4.926	-	-	-	-
Oral Surgery	5.727	0.164	2.691	4.840	1.284	0.175	0.915	0.186
Pharmacology	9.988	0.026	0.440	3.936	-	-	-	-
Small Animals	0.000	1.532	0.810	5.182	0.526	0.089	0.452	0.112
Statistics	3.899	1.048	1.419	5.232	0.367	0.093	0.330	0.041

Table D.3: *Results obtained when citation data are fitted with the Polya Aeppli model. A dash '-' indicates that the model is unsuitable.*

Subject	Estimated coefficient				Standard error			
	μ_{inter}	μ_{author}	$\mu_{countries}$	d	μ_{inter}	μ_{author}	$\mu_{countries}$	d
Applied maths	1.297	1.212	4.096	11.457	0.576	0.093	0.577	0.208
Aquatic	3.725	0.823	3.102	8.488	0.288	0.049	0.227	0.104
Archeology	0.000	0.226	2.910	8.176	0.780	0.107	0.784	0.559
Biochemistry	9.925	0.308	4.829	4.164	-	-	-	-
Biomedical	0.723	1.156	7.876	21.120	-	-	-	-
Biophysics	9.692	0.658	2.170	13.430	0.306	0.072	0.235	0.126
Care planning	1.145	1.624	−0.340	7.955	1.389	0.206	1.396	0.612
Neuroscience	10.602	0.823	3.221	14.620	-	0.064	0.581	0.124
Chemical health	4.233	0.128	2.196	6.523	1.345	0.179	1.188	0.497
Computer	8.976	0.666	−2.812	7.898	0.172	0.060	0.118	0.110
Physics	9.980	0.499	4.993	4.019	-	0.000	-	-
Developmental	5.521	1.898	1.649	12.461	0.424	0.071	0.401	0.122
Earth	3.014	1.545	2.776	11.530	0.344	0.072	0.347	0.215
Education	2.907	1.227	1.477	10.525	0.589	0.073	0.536	0.183
Electronic	10.097	0.647	1.735	15.008	0.417	0.096	0.471	0.175
Environmental	9.882	1.244	3.004	15.479	-	0.087	-	-
Inorganic	7.271	0.722	2.197	12.266	-	0.045	-	0.157
Management	5.719	1.471	1.300	14.349	-	0.210	-	0.202
Microbiology	7.304	0.734	3.564	13.005	0.695	0.078	0.333	0.099
Nuclear	5.637	0.283	1.044	12.022	0.384	0.040	0.342	0.188
Oral Surgery	6.062	0.160	2.421	6.786	1.142	0.204	0.950	0.418
Pharmacology	2.027	0.913	3.106	11.737	0.469	0.065	0.400	0.231
Small Animals	0.002	1.648	0.443	8.795	0.642	0.105	0.568	0.422
Statistics	4.485	0.912	1.169	9.593	0.473	0.132	0.406	0.153

Table D.4: *Results obtained when citation data are fitted with SVA Poisson-NB model. A dash '-' indicates that the model is unsuitable.*

Subject	Estimated coefficients							Standard errors						
	λ_{inter}	λ_{author}	$\lambda_{countries}$	μ_{inter}	μ_{author}	$\mu_{countries}$	α	λ_{inter}	λ_{author}	$\lambda_{countries}$	μ_{inter}	μ_{author}	$\mu_{countries}$	α
Applied maths	0.183	0.102	−0.035	0.694	0.124	1.125	−0.927	0.056	0.009	0.051	0.118	0.017	0.112	0.026
Aquatic	0.598	0.018	0.224	1.613	0.064	0.197	−0.561	0.025	0.007	0.015	0.038	0.007	0.026	0.022
Archeology	−0.649	0.067	0.183	0.128	−0.040	1.139	−1.124	0.115	0.016	0.115	0.392	0.053	0.407	0.094
Biochemistry	0.336	0.058	0.042	1.814	0.035	0.427	−0.614	0.028	0.003	0.008	0.070	0.008	0.043	0.031
Biomedical	−0.382	0.067	0.426	1.104	0.138	0.809	−0.778	0.058	0.007	0.044	0.101	0.011	0.080	0.028
Biophysics	1.033	0.015	0.007	1.084	0.104	0.656	−0.720	0.023	0.003	0.025	0.060	0.007	0.043	0.022
Care planning	0.069	0.145	−0.193	−0.534	0.414	0.600	−0.678	0.241	0.032	0.237	0.590	0.062	0.509	0.151
Neuroscience	0.928	0.035	0.061	1.774	0.080	0.345	−0.502	0.020	0.002	0.009	0.045	0.006	0.028	0.023
Chemical health	0.486	0.037	0.264	0.982	−0.002	0.560	−0.960	0.201	0.024	0.164	0.431	0.054	0.310	0.179
Computer	0.011	0.028	0.300	1.208	0.087	0.682	−0.975	0.070	0.014	0.063	0.133	0.023	0.117	0.034
Physics	0.330	0.047	0.122	1.358	0.104	0.626	−1.099	0.040	0.007	0.033	0.088	0.012	0.074	0.025
Developmental	0.921	0.018	−0.176	1.041	0.143	0.835	−0.650	0.045	0.001	0.041	0.083	0.010	0.073	0.023
Earth	0.309	0.089	−0.056	1.406	0.107	0.415	−0.557	0.024	0.005	0.026	0.054	0.010	0.045	0.025
Education	0.728	0.071	−0.379	4.161	−0.173	−0.582	−1.084	0.100	0.008	0.098	-	-	0.052	-
Electronic	0.617	0.028	0.129	2.194	0.050	0.207	−0.793	0.044	0.007	0.035	0.071	0.011	0.056	0.026
Environmental	0.731	0.079	−0.101	2.342	0.031	0.287	−0.385	0.020	0.006	0.026	0.041	0.007	0.031	0.018
Inorganic	0.426	0.099	−0.003	1.999	0.056	0.134	−0.772	0.042	0.006	0.036	0.053	0.008	0.042	0.020
Management	0.214	0.128	0.024	0.383	0.156	1.540	−1.045	0.101	0.022	0.093	0.450	0.046	0.443	0.056
Microbiology	0.706	0.000	0.124	1.234	0.070	0.639	−0.544	0.042	0.005	0.029	0.048	0.005	0.033	0.022
Nuclear	0.112	0.024	0.153	2.512	−0.034	0.035	−0.897	0.023	0.005	0.029	0.052	0.004	0.037	0.028
Oral Surgery	0.922	0.011	0.042	1.426	0.024	0.355	−0.454	0.223	0.025	0.161	0.266	0.038	0.175	0.115
Pharmacology	−0.079	0.071	0.226	1.874	0.038	0.242	−0.578	0.047	0.006	0.030	0.073	0.009	0.049	0.031
Small Animals	−0.564	0.168	0.380	4.834	−0.351	−0.472	−1.017	0.067	0.010	0.048	-	-	0.105	-
Statistics	0.797	0.040	−0.246	−0.512	0.202	1.785	−1.254	0.059	0.014	0.056	0.276	0.031	0.261	0.043

Table D.5: *Results obtained when citation data are fitted with SVA NB-Poisson model. A dash '-' indicates that the model is unsuitable. This model is unsuitable for the subject 'Computer Graphics and Computer Aided Design'.*

Subject	Estimated coefficients							Standard error						
	μ_{inter}	μ_{author}	$\mu_{countries}$	α	λ_{inter}	λ_{author}	$\lambda_{countries}$	μ_{inter}	μ_{author}	$\mu_{countries}$	α	λ_{inter}	λ_{author}	$\lambda_{countries}$
Applied maths	1.253	0.176	0.382	−0.583	2.033	−3.005	−5.295	0.059	0.013	0.052	0.019	332.913	30.444	331.613
Aquatic	1.807	0.072	0.206	0.081	−3.338	−2.466	3.236	0.027	0.005	0.019	0.016	0.564	-	0.647
Archeology	0.264	0.219	0.154	−1.022	−5.515	−5.196	−0.303	0.204	0.045	0.209	0.058	2097.152	2117.880	288.980
Biochemistry	1.624	0.089	0.327	−0.391	2.345	−5.023	−0.004	0.057	0.007	0.036	0.025	-	15.717	-
Biomedical	1.865	0.097	0.311	−0.814	0.283	−4.154	2.697	0.072	0.010	0.053	0.021	-	-	0.612
Biophysics	2.342	0.048	0.134	−0.094	0.308	−1.883	−0.118	0.035	0.005	0.023	0.015	-	-	-
Care planning	0.171	0.394	0.043	−0.515	0.233	−4.450	2.633	0.385	0.048	0.333	0.109	-	20.608	19.024
Neuroscience	2.486	0.035	0.177	0.041	0.312	0.043	−0.843	0.032	0.004	0.019	0.018	-	0.004	-
Chemical health	1.085	0.017	0.664	0.023	−0.496	0.154	−1.000	0.247	0.032	0.194	0.099	-	0.165	-
Physics	1.685	0.107	0.373	−0.613	−1.032	−0.452	−0.631	0.059	0.009	0.048	0.019	7.876	0.541	7.087
Developmental	1.905	0.122	0.206	−0.258	−1.743	0.011	0.440	0.046	0.008	0.038	0.018	0.169	-	0.036
Earth	1.835	0.111	0.147	−0.333	−0.948	0.214	−2.275	0.039	0.009	0.033	0.020	-	-	-
Education	1.310	0.177	0.215	−0.720	−1.213	−1.447	−5.932	0.093	0.013	0.088	0.019	597.197	-	603.389
Electronic	2.290	0.061	0.113	−0.315	1.543	−1.873	−0.380	0.050	0.008	0.039	0.020	-	1.088	-
Environmental	2.503	0.060	0.127	−0.109	−1.024	0.120	−1.098	0.034	0.006	0.025	0.016	-	0.112	-
Inorganic	2.212	0.039	0.151	−0.208	−5.890	−1.879	2.318	0.039	0.006	0.030	0.015	-	-	-
Management	1.832	0.189	0.027	−0.678	−2.456	0.430	−1.283	0.136	0.034	0.128	0.035	0.958	0.436	-
Microbiology	2.219	0.057	0.149	−0.085	−1.400	0.087	0.101	0.031	0.004	0.019	0.018	0.200	0.008	0.026
Nuclear	1.878	0.019	0.108	−0.768	4.469	−5.547	−4.332	0.047	0.005	0.041	0.021	1482.910	-	1480.885
Oral Surgery	1.812	0.019	0.276	0.262	−1.103	−0.023	0.179	0.160	0.025	0.104	0.087	-	0.071	-
Pharmacology	1.550	0.076	0.294	−0.441	−2.351	−4.727	2.028	0.062	0.008	0.044	0.025	-	-	-
Small Animals	0.655	0.232	0.202	−0.655	−5.468	−4.176	2.912	0.138	0.023	0.101	0.054	-	-	-
Statistics	1.492	0.163	0.151	−0.517	−3.340	0.060	0.687	0.066	0.020	0.056	0.022	0.420	0.051	0.317

Table D.6: *Results obtained when citation data are fitted with SVA NB-NB model. A dash '-' indicates that the model is unsuitable.*

Subject	Estimated coefficients								Standard error							
	μ_{inter1}	μ_{auth1}	μ_{coun1}	α_1	μ_{inter2}	μ_{auth2}	μ_{coun2}	α_2	μ_{inter1}	μ_{auth1}	μ_{coun1}	α_1	μ_{inter2}	μ_{auth2}	μ_{coun2}	α_2
Applied maths	1.20	0.18	0.32	−0.47	0.11	−0.53	3.73	−6.27	0.06	0.01	0.05	0.02	9.20	0.26	7.64	-
Aquatic	0.97	0.11	0.37	0.42	2.15	−0.02	−0.44	−1.56	0.05	0.01	0.03	0.04	0.15	0.02	0.14	0.10
Archeology	0.21	0.19	0.19	−0.99	5.76	−1.22	−1.76	−6.40	0.20	0.04	0.20	0.06	5.96	1.53	6.69	-
Biochemistry	0.67	0.05	−0.03	2.04	0.83	0.19	0.32	−0.76	0.03	0.01	0.02	-	0.07	0.01	0.05	0.03
Biomedical	−0.75	0.08	0.87	1.55	2.50	0.09	−0.13	−0.74	0.06	0.01	0.02	-	0.07	0.01	0.05	0.02
Biophysics	2.22	0.05	−0.90	1.21	1.12	0.15	0.61	−0.58	-	0.01	0.02	-	0.02	0.01	0.03	-
Care planning	−0.33	0.17	0.02	2.75	0.04	0.36	0.25	−0.72	0.21	0.03	0.22	-	0.51	0.06	0.44	0.10
Neuroscience	2.37	0.01	−0.04	0.74	2.47	−0.02	0.11	−1.62	0.15	0.01	0.09	0.18	0.24	0.03	0.08	0.80
Chemical health	2.44	−0.10	0.04	0.06	0.98	0.33	−4.33	4.51	0.25	0.03	0.15	0.11	-	0.13	-	16.88
Computer	1.06	0.00	0.90	−0.67	2.16	−1.92	5.04	−5.68	0.11	0.02	0.10	0.03	5.70	0.13	3.85	-
Physics	0.56	−0.03	0.27	3.49	1.38	−0.02	0.94	−0.94	0.04	0.01	0.03	-	0.08	0.01	0.07	0.02
Developmental	1.52	0.12	0.31	−0.08	2.30	0.14	−1.36	−2.54	0.07	0.01	0.04	0.02	-	0.03	-	0.15
Earth	1.57	0.20	−0.02	−0.21	2.03	−1.97	1.95	−2.56	0.03	0.01	0.03	0.02	-	-	-	-
Education	0.95	0.27	0.31	−0.64	9.38	−4.87	5.48	−6.88	0.10	0.01	0.10	0.02	11.94	0.80	11.46	-
Electronic	0.67	0.05	−0.02	2.50	2.27	0.03	0.20	−0.70	0.06	0.01	0.04	0.23	0.07	0.01	0.05	0.03
Environmental	2.48	0.06	0.12	−0.02	6.39	−0.15	−1.48	−7.07	0.03	0.01	0.02	0.02	3.87	1.05	3.82	-
Inorganic	0.41	0.20	0.30	0.40	4.21	−0.18	−1.15	−1.37	0.05	0.01	0.03	0.02	-	0.01	-	0.03
Management	0.03	0.16	0.26	1.81	0.67	0.01	1.54	−1.12	0.10	0.02	0.10	-	0.35	0.05	0.35	0.03
Microbiology	2.19	0.05	0.15	−0.12	−0.15	−0.14	0.66	−0.31	0.03	0.00	0.02	-	-	-	-	0.20
Nuclear	1.68	0.03	0.15	−0.68	2.14	−0.34	2.46	−6.46	0.05	0.01	0.05	0.02	11.98	0.26	9.85	-
Oral Surgery	1.43	0.04	0.47	0.45	5.11	−1.30	1.08	−3.78	0.18	0.03	0.13	0.10	2.66	-	2.02	0.07
Pharmacology	0.80	0.06	−0.37	3.83	2.37	0.00	0.16	−0.81	0.06	0.01	0.05	-	0.09	0.01	0.05	0.03
Small Animals	0.08	0.40	−0.18	−0.19	2.45	−0.80	1.23	−1.09	0.15	0.02	0.11	0.06	0.39	-	0.33	0.15
Statistics	1.18	0.16	0.04	−0.05	−1.15	0.24	1.71	−3.12	0.06	0.02	0.06	0.02	-	0.08	-	-

Table D.7: *Results obtained when citation data are fitted with SVB Poisson-NB model. A dash '-' indicates that the model is unsuitable. This table excludes Care Planning, Cellular and Molecular Neuroscience, Condensed Matter Physics, Nuclear Energy and Engineering, as the model is unsuitable for these subjects.*

Subject	Estimated coefficients							Standard error						
	λ_{inter}	λ_{author}	$\lambda_{countries}$	μ_{inter}	μ_{author}	$\mu_{countries}$	α	λ_{inter}	λ_{author}	$\lambda_{countries}$	μ_{inter}	μ_{author}	$\mu_{countries}$	α
Applied maths	−0.254	−13.141	6.653	1.257	0.176	0.379	−0.588	-	-	5.428	0.059	0.013	0.052	0.019
Aquatic	−2.855	0.047	0.451	1.761	0.078	0.206	0.041	2.933	0.028	0.245	0.042	0.006	0.020	0.101
Archeology	1.749	−11.295	0.841	0.256	0.217	0.164	−1.024	-	-	-	0.204	0.044	0.210	0.058
Biochemistry	−1.950	−2.325	2.620	1.697	0.077	0.331	−0.390	0.346	0.691	0.747	0.057	0.007	0.035	0.026
Biomedical	0.377	−11.777	2.837	1.859	0.097	0.312	−0.826	1048.607	1047.885	-	0.073	0.010	0.053	0.021
Biophysics	0.237	−4.066	0.301	2.381	0.041	0.130	−0.084	20.675	21.206	-	0.034	0.005	0.023	0.015
Chemical health	2.183	−2.306	−0.932	1.496	0.008	0.397	0.015	-	-	-	0.239	0.031	0.178	0.105
Computer	2.838	−5.098	−4.334	1.469	0.075	0.438	−0.690	-	-	-	0.093	0.017	0.080	0.025
Developmental	−1.365	−2.873	−1.636	2.026	0.133	0.091	−0.224	63.319	-	64.061	0.042	0.008	0.034	0.017
Earth	3.667	−2.180	−8.233	1.836	0.115	0.139	−0.329	-	15.662	-	0.039	0.009	0.033	0.020
Education	−8.009	−1.619	3.238	0.983	0.164	0.558	−0.703	2.806	-	0.855	0.116	0.012	0.112	0.019
Electronic	4.865	0.096	−6.317	2.036	0.063	0.288	−0.317	-	0.049	-	0.056	0.008	0.044	0.027
Environmental	−3.638	−5.925	−0.245	2.510	0.062	0.121	−0.088	48.299	741.455	-	0.033	0.006	0.024	0.015
Inorganic	1.499	−4.471	0.410	2.237	0.039	0.137	−0.208	8.908	8.208	3.085	0.038	0.006	0.030	0.015
Management	−1.109	0.129	−0.342	1.606	0.127	0.393	−0.808	-	0.134	-	0.168	0.036	0.163	0.046
Microbiology	1.678	−7.474	4.397	2.264	0.047	0.179	−0.015	-	-	0.444	0.029	0.004	0.019	0.016
Oral Surgery	1.645	−2.564	−0.287	1.837	0.020	0.275	0.310	25.468	-	24.862	0.170	0.024	0.114	0.081
Pharmacology	−2.235	0.237	−2.582	1.564	0.075	0.293	−0.454	-	0.028	-	0.063	0.008	0.044	0.026
Small Animals	−0.507	0.014	−1.908	0.605	0.223	0.253	−0.603	-	-	-	0.136	0.022	0.100	0.052
Statistics	1.568	−4.832	1.099	1.525	0.149	0.157	−0.481	-	-	1.389	0.065	0.020	0.054	0.023

Table D.8: *Results obtained when citation data are fitted with SVB NB-NB model. A dash '-' indicates that the model is unsuitable.*

Subject	Estimated coefficients								Standard error							
	μ_{inter1}	μ_{auth1}	μ_{coun1}	α_1	μ_{inter2}	μ_{auth2}	μ_{coun2}	α_2	μ_{inter1}	μ_{auth1}	μ_{coun1}	α_1	μ_{inter2}	μ_{auth2}	μ_{coun2}	α_2
Applied maths	−1.92	−0.34	1.17	2.50	0.94	0.19	0.63	−0.68	0.19	-	-	-	0.07	0.01	0.06	-
Aquatic	9.90	−8.93	−8.81	2.88	1.81	0.07	0.20	0.09	-	741.47	1048.59	0.00	0.03	0.01	0.02	0.02
Archeology	0.81	−0.04	0.30	−1.16	10.16	−0.81	3.87	−9.59	0.19	0.04	0.19	0.06	2151.31	155.14	2047.57	-
Biochemistry	4.26	−8.43	3.56	12.88	2.07	0.02	0.34	−0.40	7.37	15.24	6.38	-	0.06	0.01	0.03	0.03
Biomedical	3.55	−9.10	2.26	2.47	2.36	0.05	0.16	−0.83	270.07	270.25	-	-	0.07	0.01	0.05	0.02
Biophysics	7.58	−14.34	2.52	1.30	2.37	0.05	0.11	−0.09	741.46	741.46	0.31	-	0.03	0.00	0.02	0.02
Care planning	1.69	1.29	−15.59	3.72	2.03	0.25	−1.12	−0.66	-	-	-	-	-	-	-	-
Neuroscience	8.64	−10.21	−1.80	−2.46	2.60	0.04	0.13	0.13	146.79	152.94	-	-	0.03	0.00	0.02	0.02
Chemical health	7.73	−12.14	−3.48	3.11	1.32	0.01	0.51	0.11	-	-	-	-	-	-	-	-
Computer	5.67	−14.36	3.13	5.03	1.47	0.07	0.44	−0.70	908.13	1738.87	1048.58	-	0.09	0.02	0.08	0.03
Physics	3.57	0.02	−3.75	1.99	1.60	0.08	0.45	−0.97	-	0.03	-	-	0.07	0.01	0.06	0.03
Developmental	−0.17	−6.55	−1.99	−3.65	2.02	0.13	0.10	−0.23	134.63	-	290.05	-	0.04	0.01	0.03	0.02
Earth	15.12	−21.69	−2.80	−2.70	1.85	0.11	0.14	−0.33	-	-	-	-	-	-	-	-
Education	−3.71	−11.00	−0.54	13.07	1.31	0.18	0.22	−0.72	-	-	-	-	-	-	-	-
Electronic	1.65	−0.02	−1.63	4.33	1.87	0.06	0.38	−0.65	-	0.03	-	-	0.07	0.01	0.05	0.03
Environmental	−2.43	−3.11	0.29	2.11	2.51	0.06	0.12	−0.09	-	21.72	-	-	0.03	0.01	0.02	0.02
Inorganic	5.56	−19.81	4.56	2.32	2.21	0.04	0.15	−0.21	-	-	-	-	-	-	-	-
Management	−2.97	−8.90	3.21	10.57	1.93	0.20	−0.05	−0.69	-	-	-	-	-	-	-	-
Microbiology	10.78	−18.13	6.43	−4.85	2.24	0.05	0.18	−0.03	121.38	264.20	137.13	-	0.03	0.00	0.02	0.02
Nuclear	9.67	−18.34	−2.42	2.12	1.90	0.02	0.12	−0.77	-	-	-	-	-	-	-	-
Oral Surgery	3.51	−0.38	−1.06	−0.48	0.65	0.13	0.65	−0.03	-	-	0.88	-	-	0.03	-	-
Pharmacology	9.63	−11.85	0.44	0.63	1.73	0.09	0.20	−0.56	-	149.01	-	-	0.07	0.01	0.05	0.03
Small Animals	5.23	−17.56	8.59	2.24	1.06	0.17	0.10	−0.67	741.46	741.46	0.42	-	0.14	0.02	0.10	0.05
Statistics	1.54	0.14	0.16	−0.47	6.80	−13.20	−1.35	3.58	0.06	0.02	0.05	0.02	-	-	70.16	69.24

Appendix E

Randomised quantile residual plots for citation analysis with no covariates

The randomised quantile residual plots for the negative binomial, discretised log-normal and proposed variant models fitted in Section 5.3 are presented here.

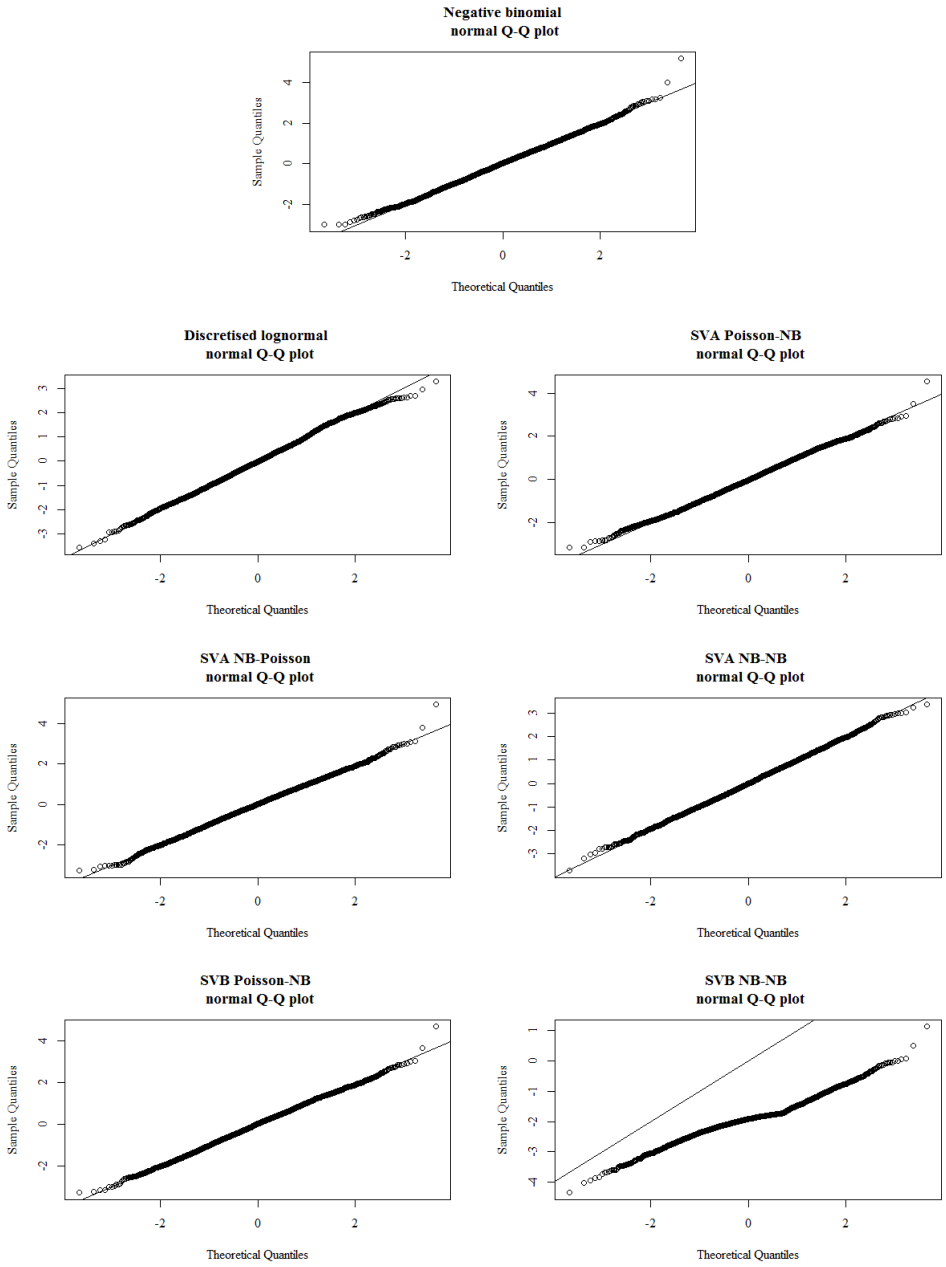

Figure E.1: *Randomised quantile residual plots of models for Visual.*

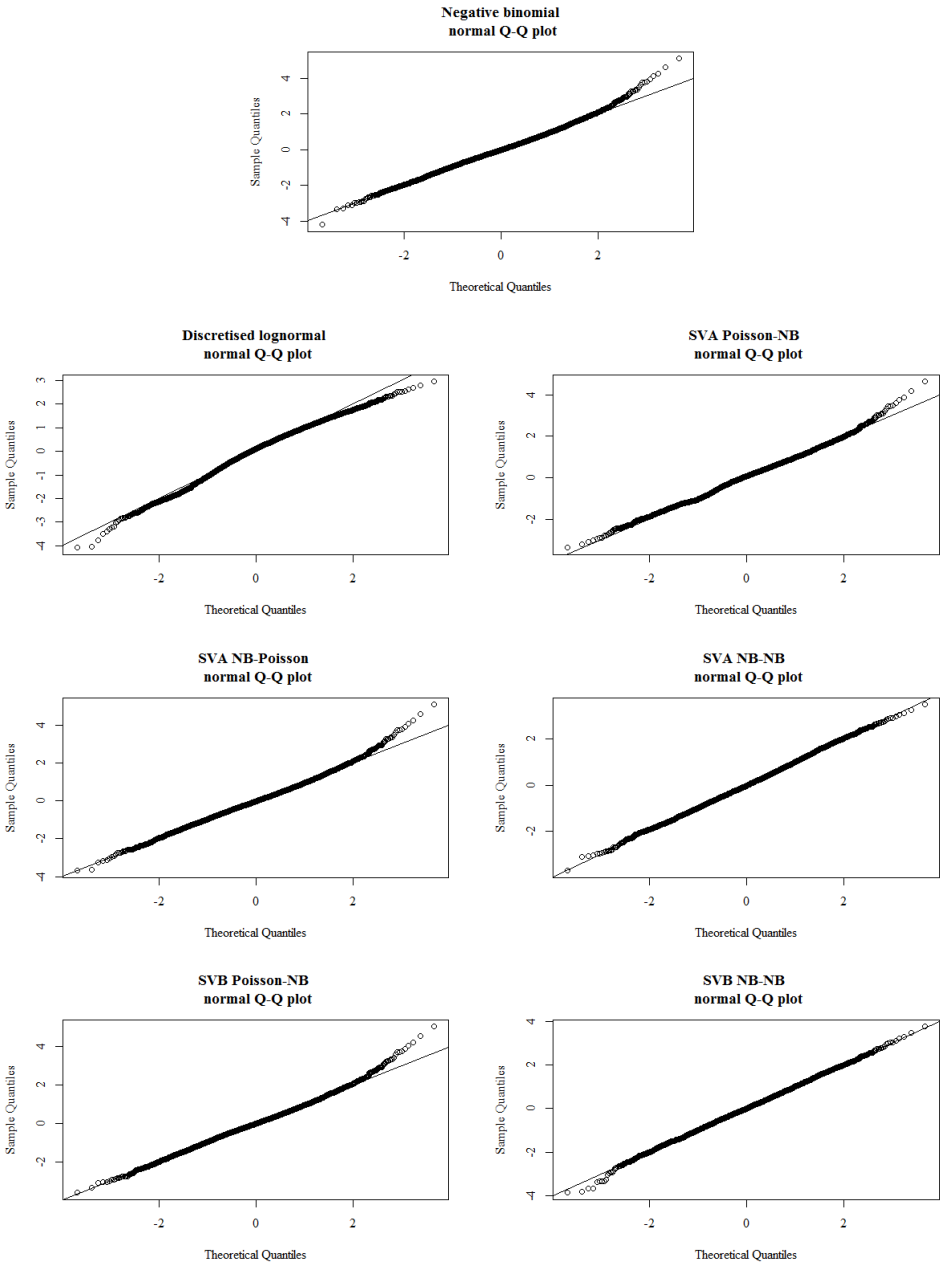

Figure E.2: *Randomised quantile residual plots of models for Soil.*

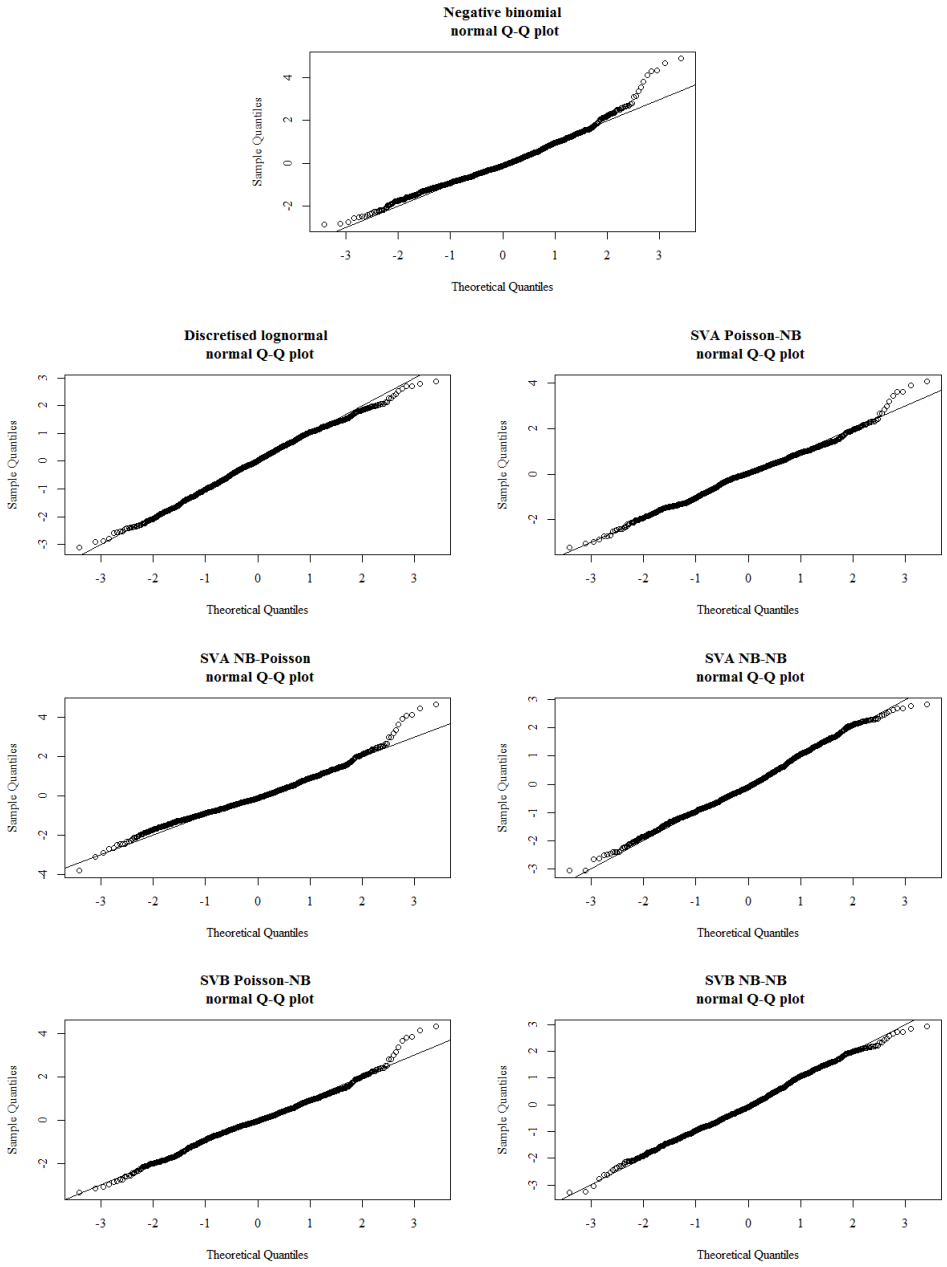

Figure E.3: *Randomised quantile residual plots of models for Marketing.*

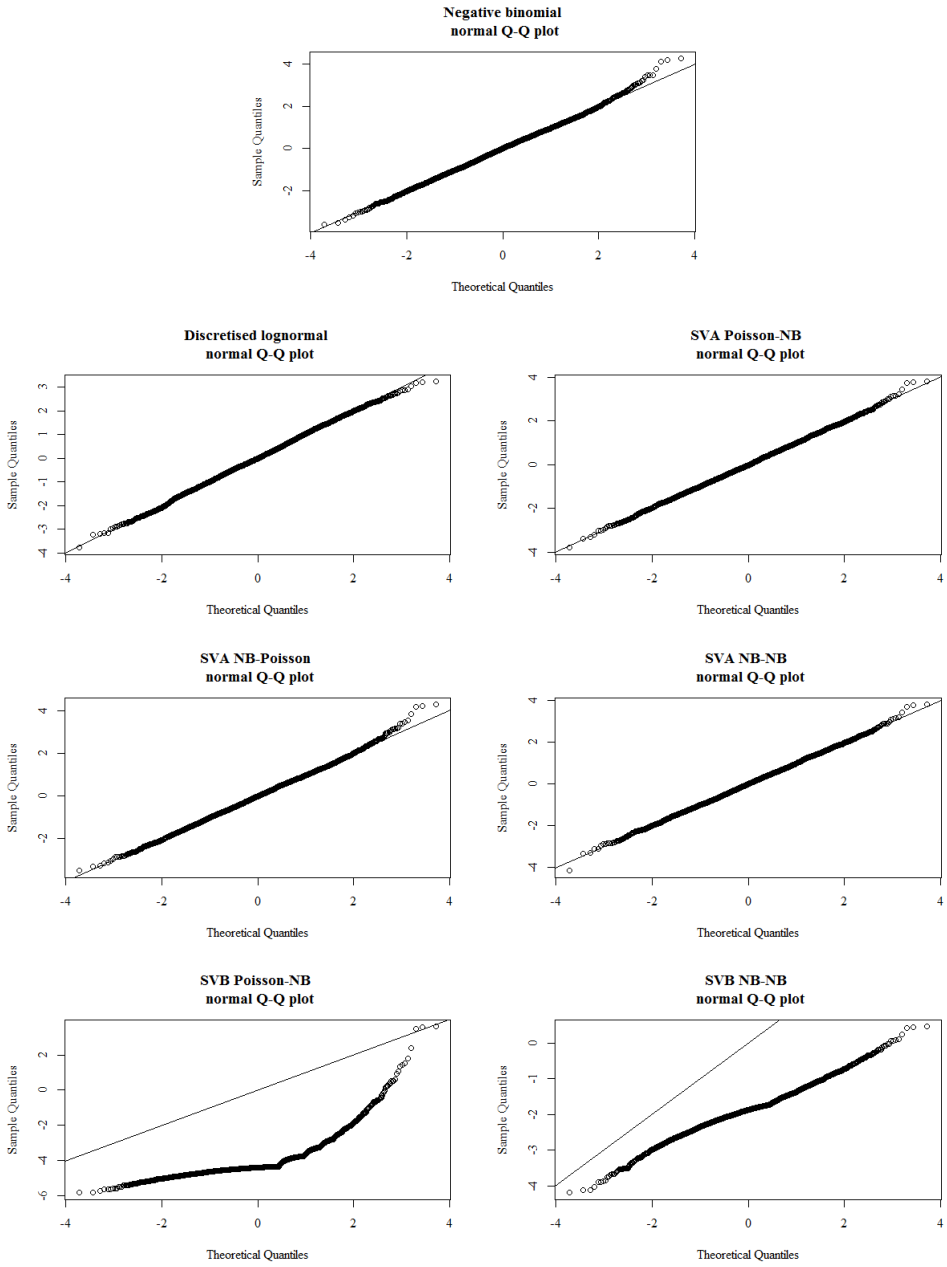

Figure E.4: *Randomised quantile residual plots of models for Literature.*

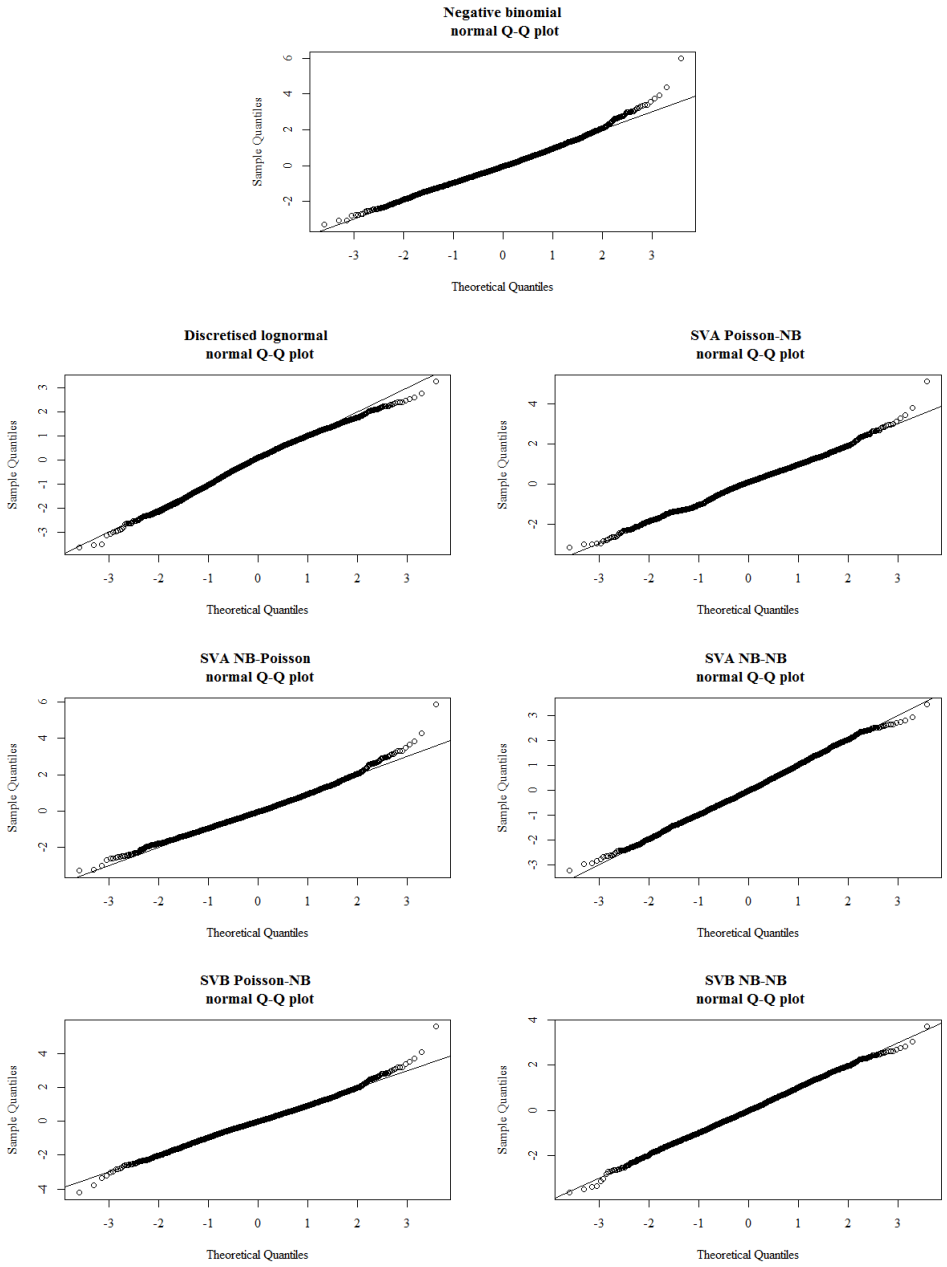

Figure E.5: *Randomised quantile residual plots of models for Horticulture.*

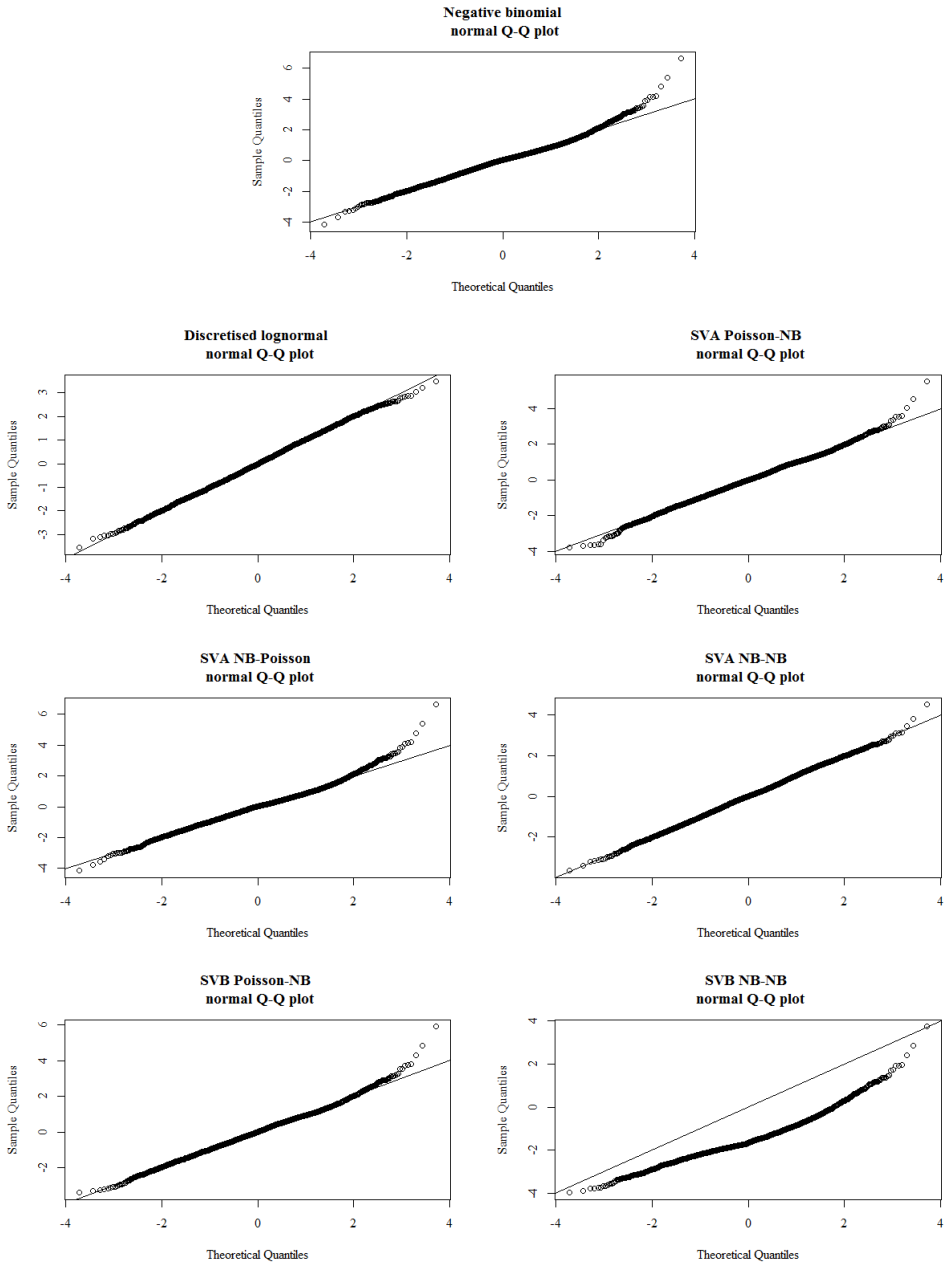

Figure E.6: *Randomised quantile residual plots of models for History.*

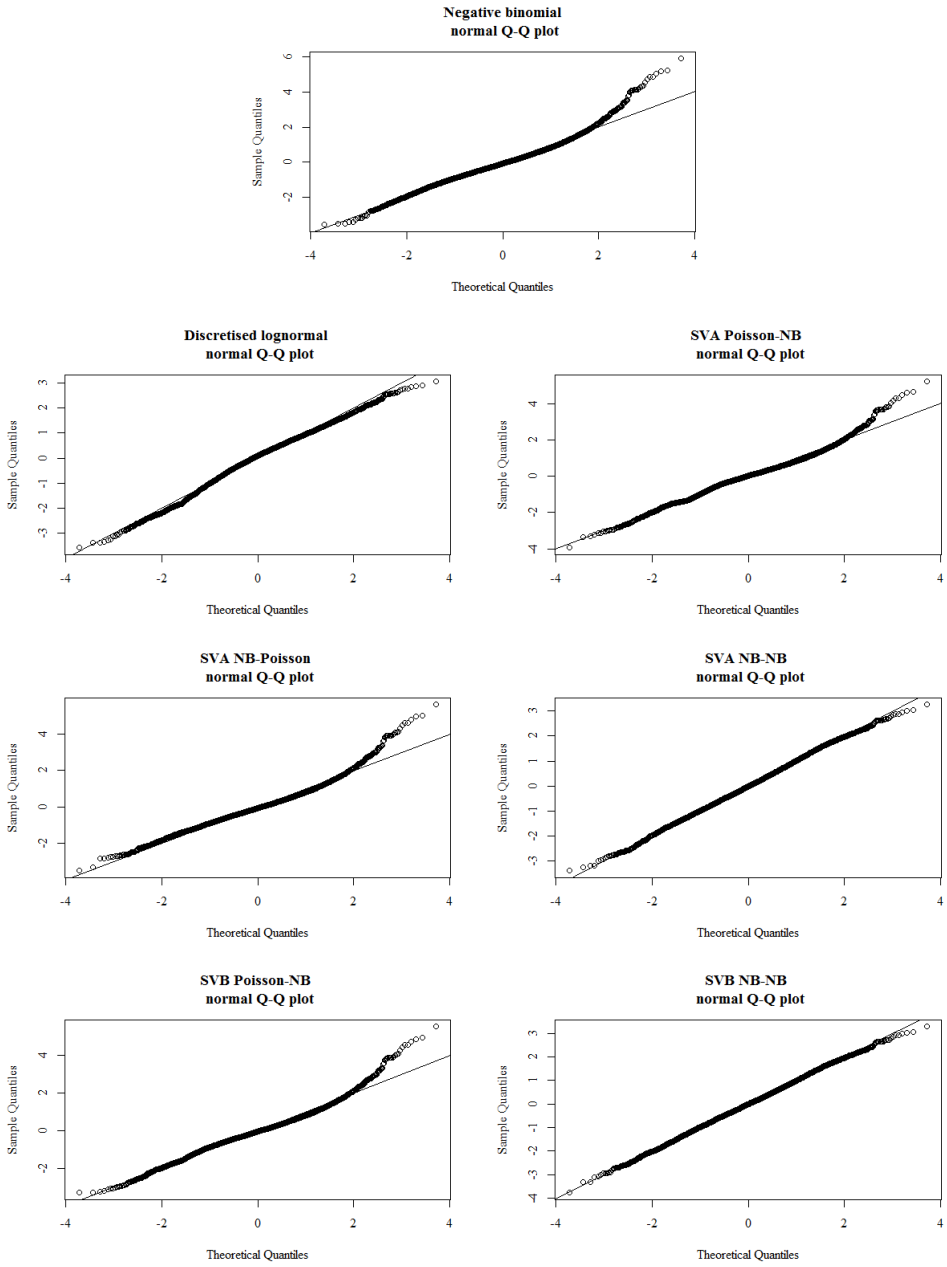

Figure E.7: *Randomised quantile residual plots of models for Genetics.*

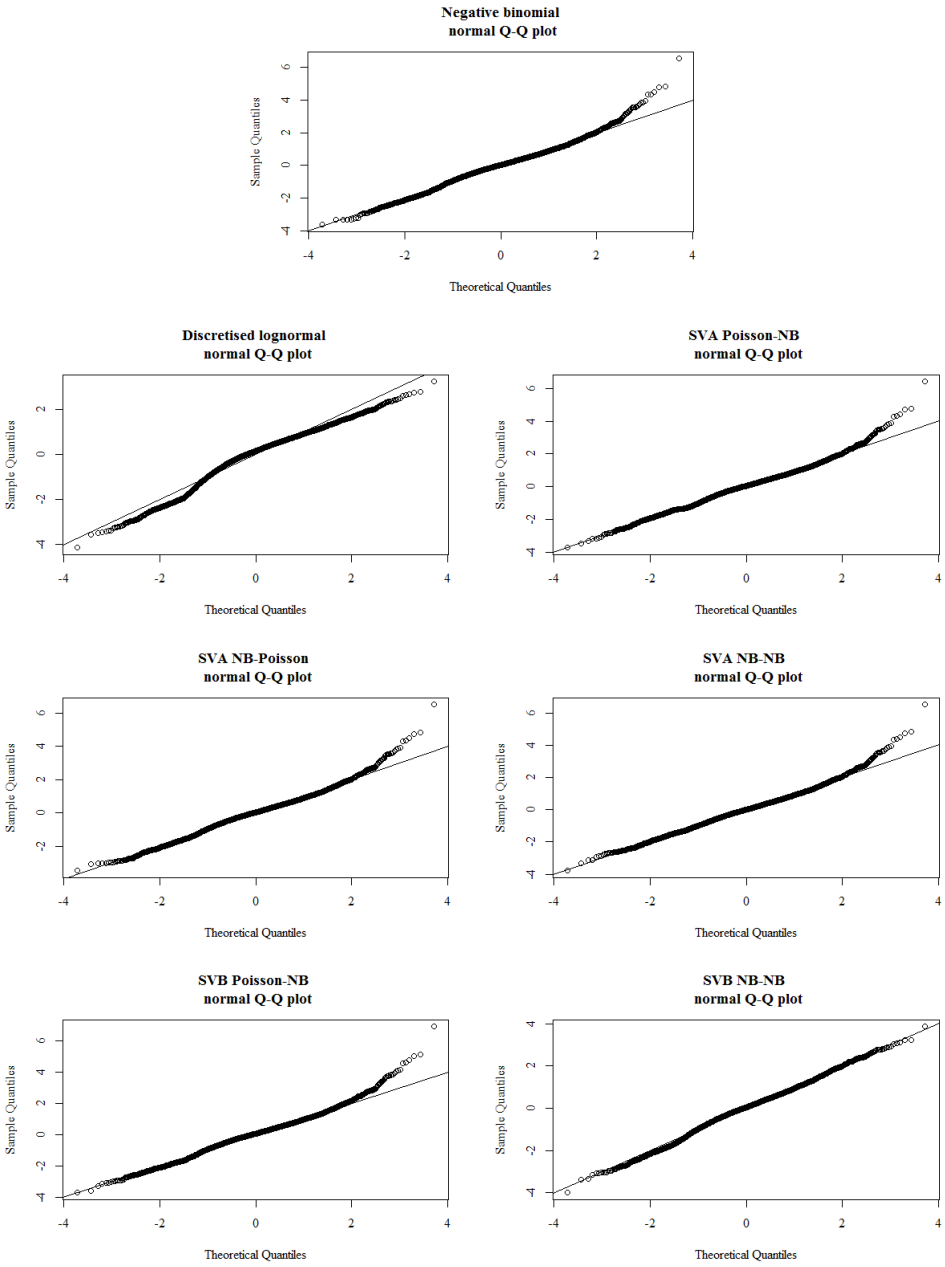

Figure E.8: *Randomised quantile residual plots of models for Ecology.*

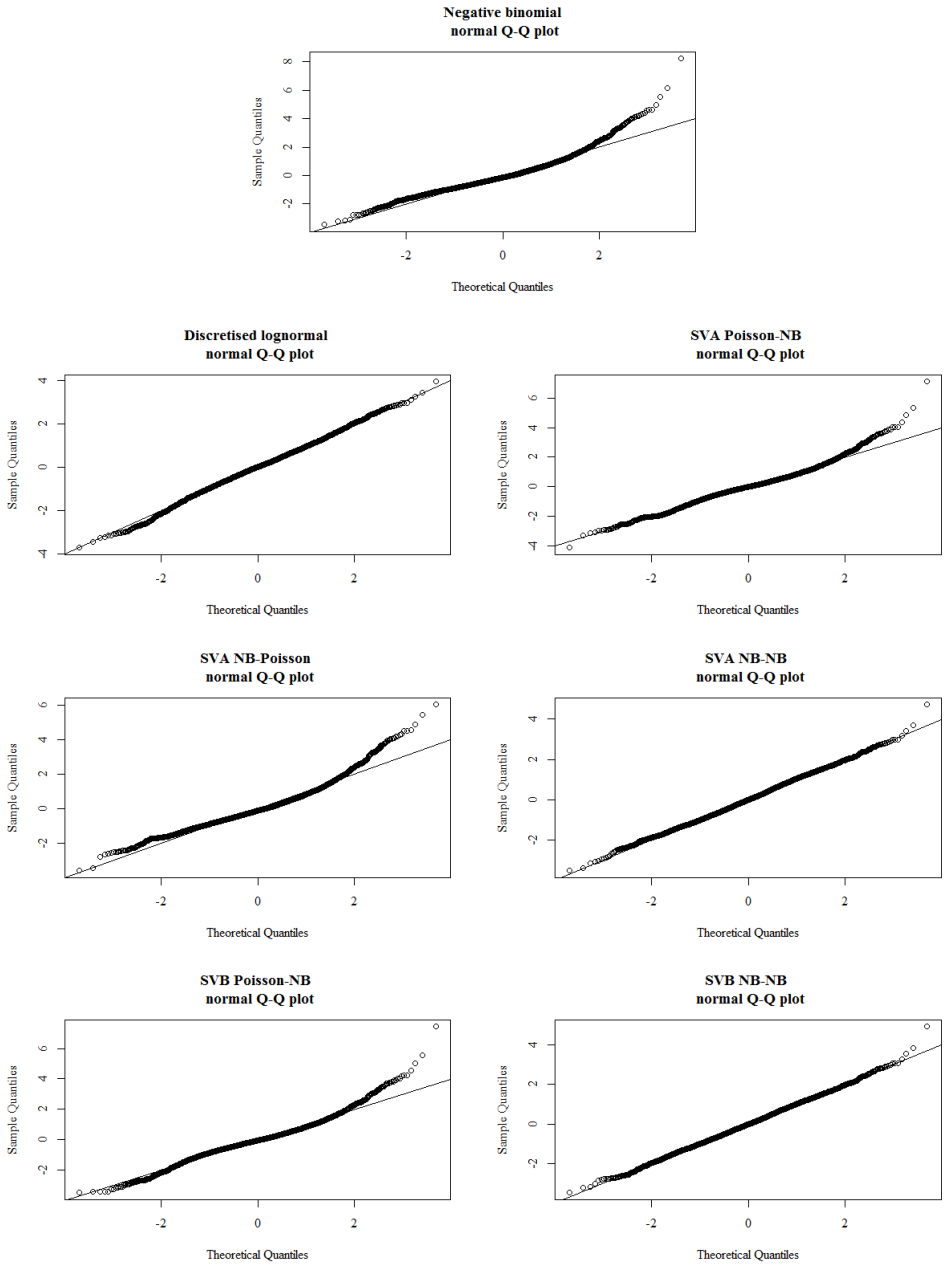

Figure E.9: *Randomised quantile residual plots of models for Developmental.*

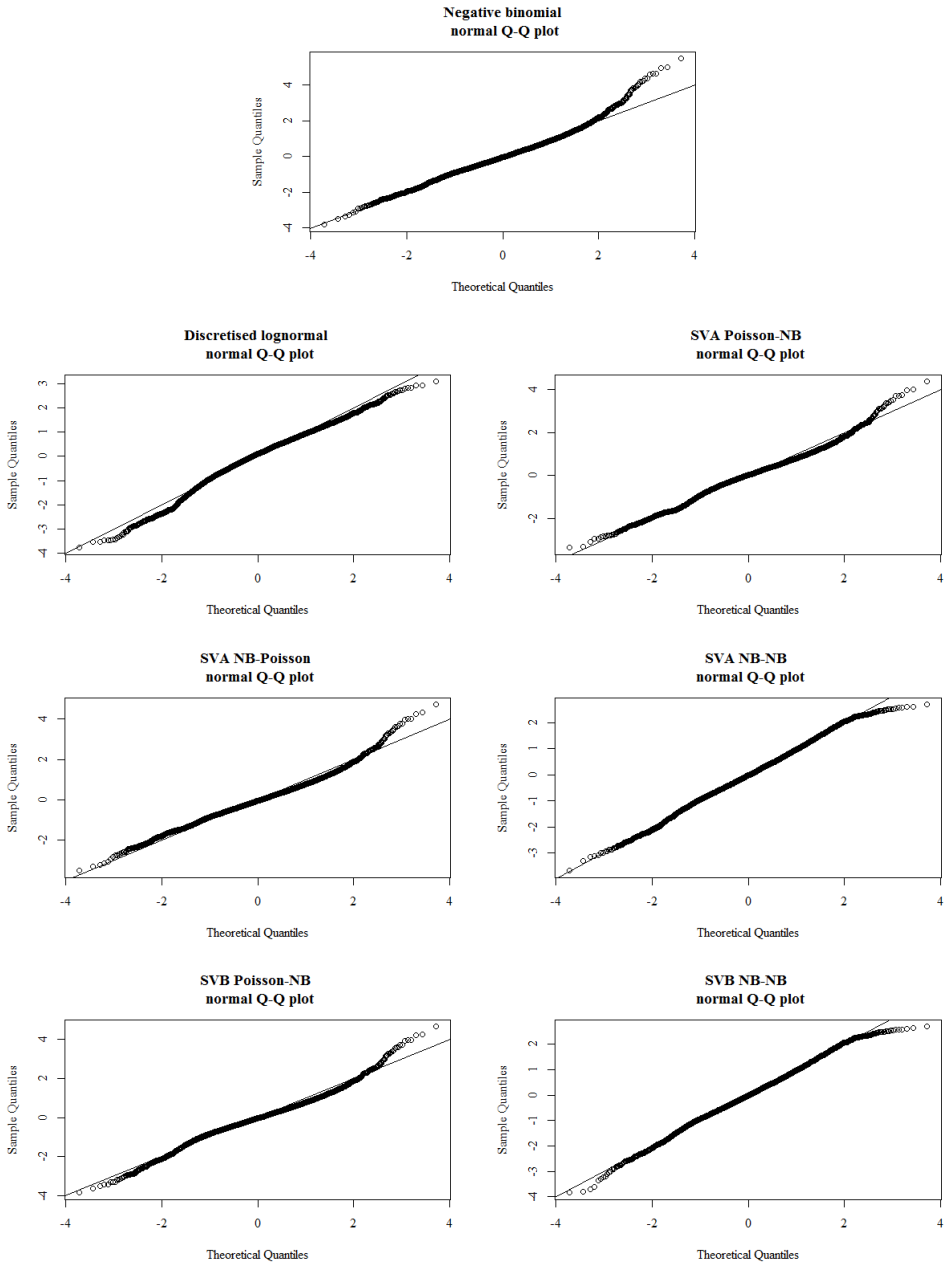

Figure E.10: *Randomised quantile residual plots of models for Biochemistry.*

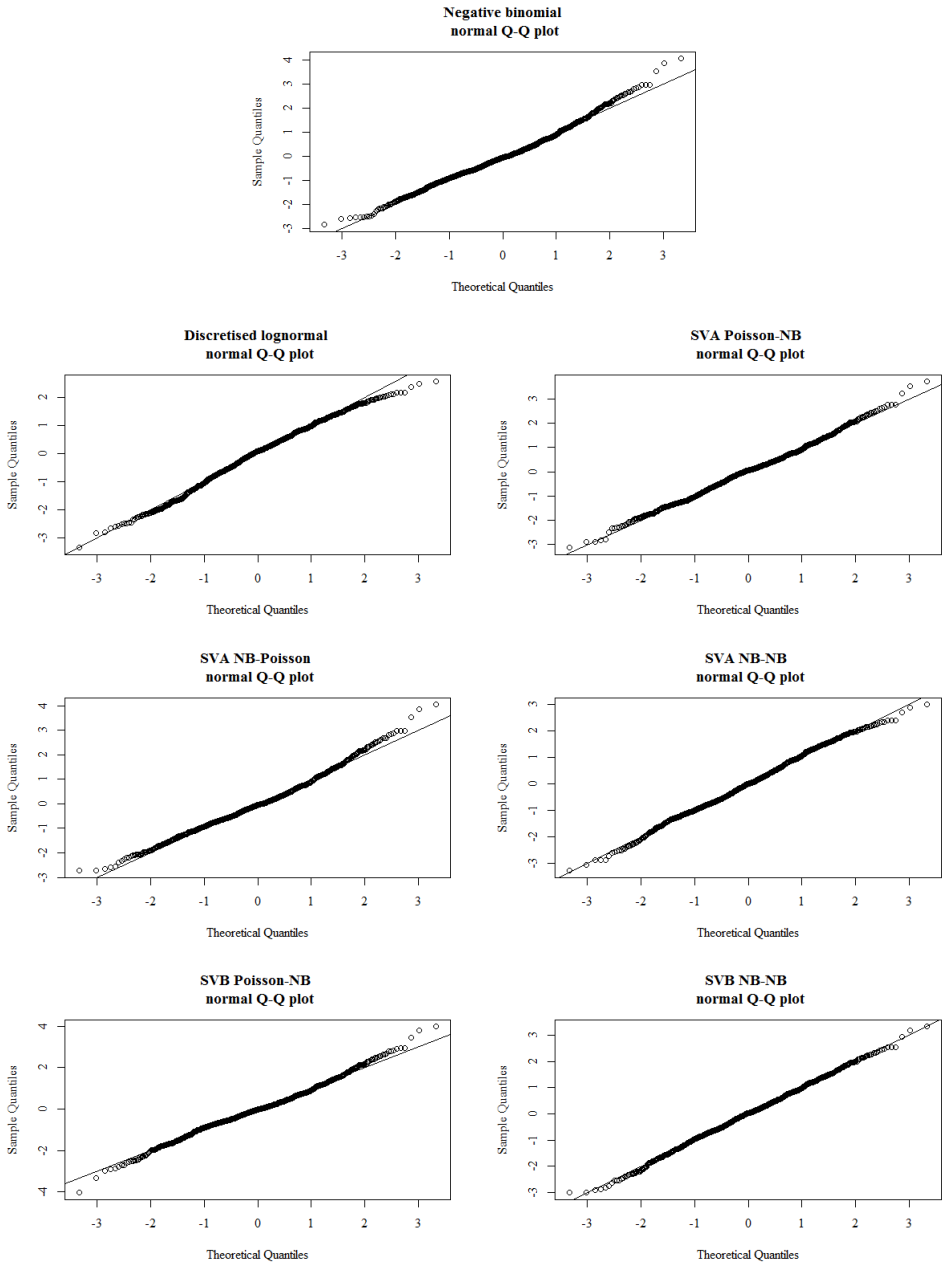

Figure E.11: *Randomised quantile residual plots of models for Accounting.*

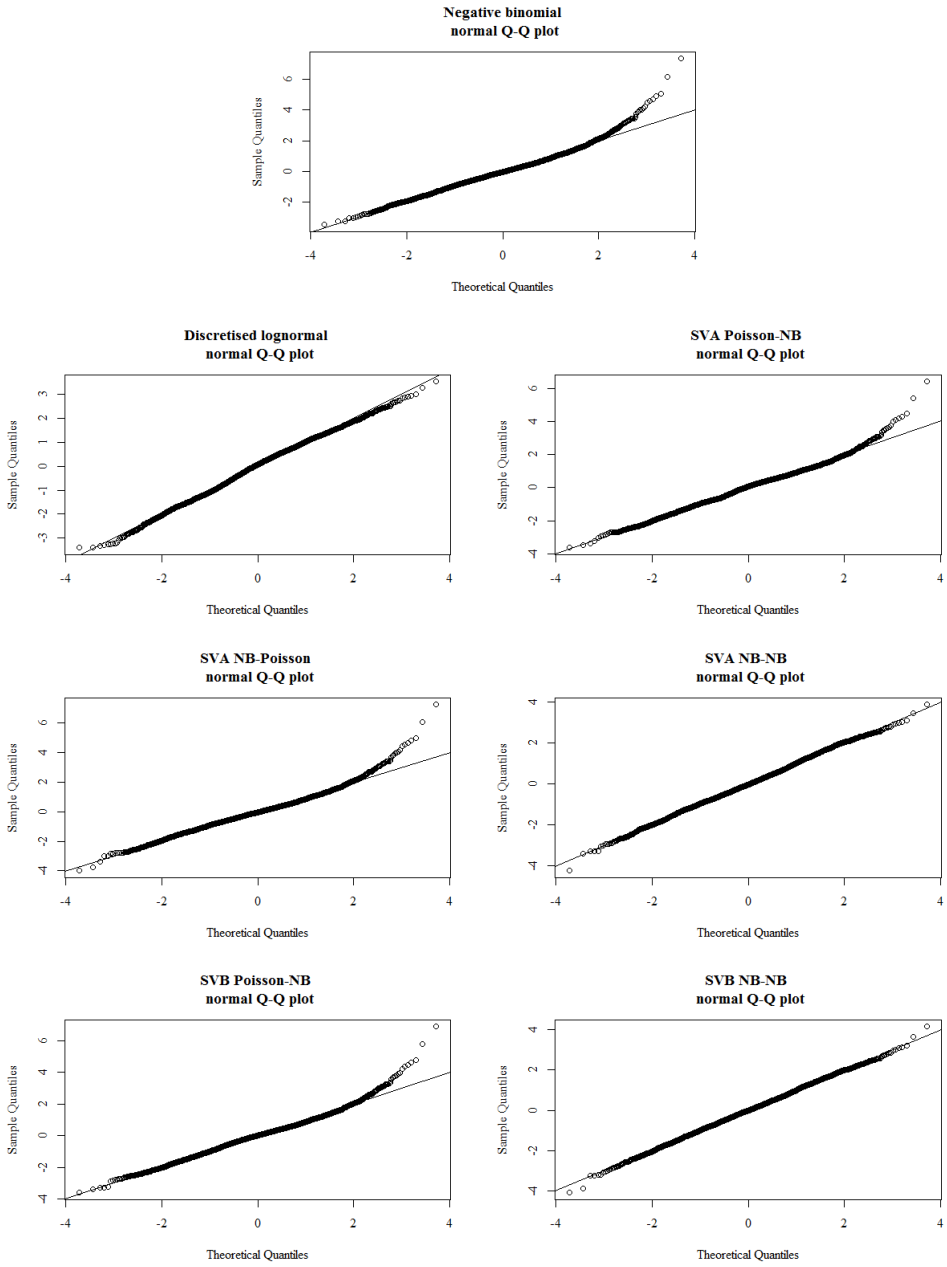

Figure E.12: *Randomised quantile residual plots of models for AppliedMaths.*

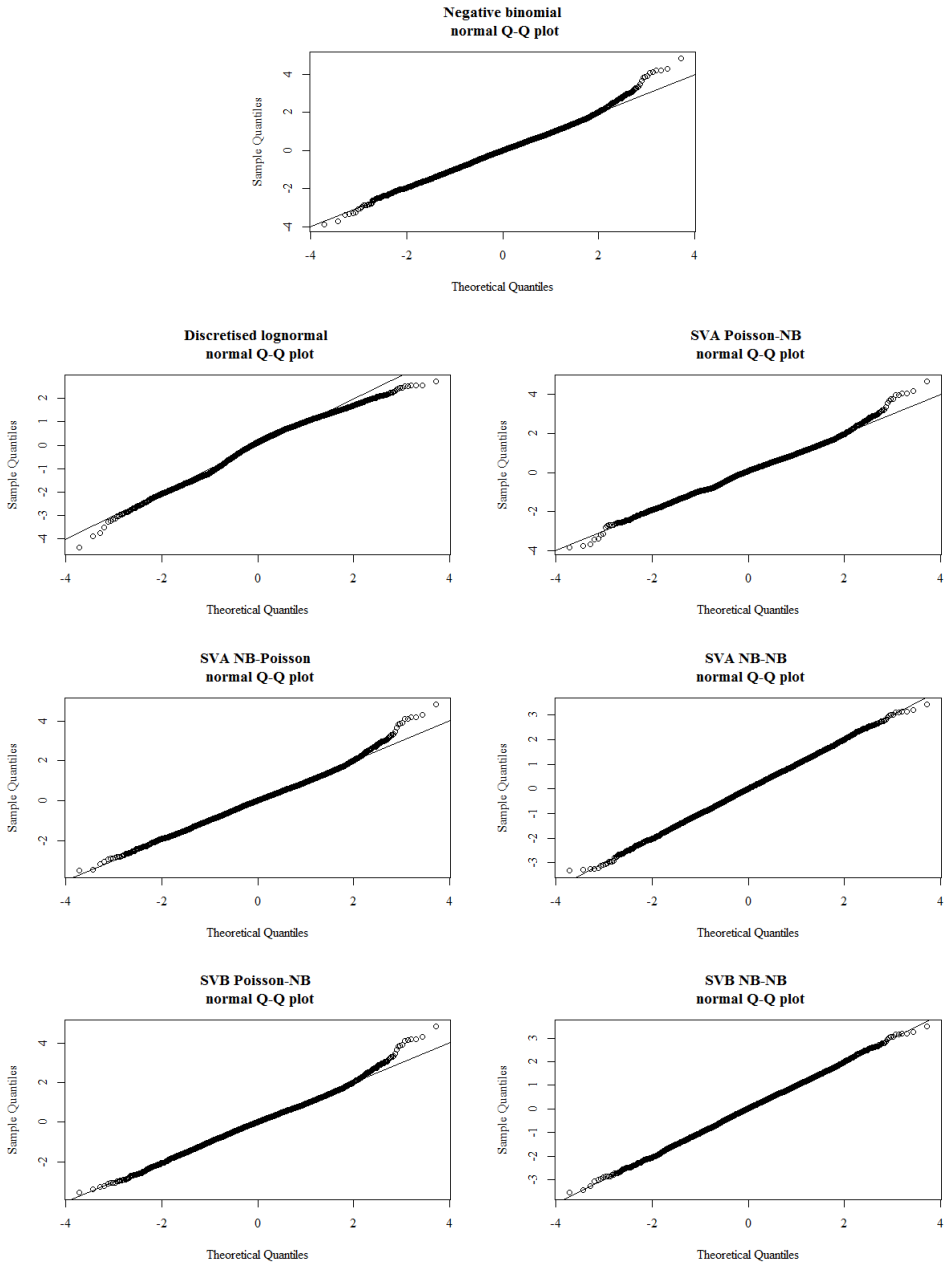

Figure E.13: *Randomised quantile residual plots of models for Urology.*

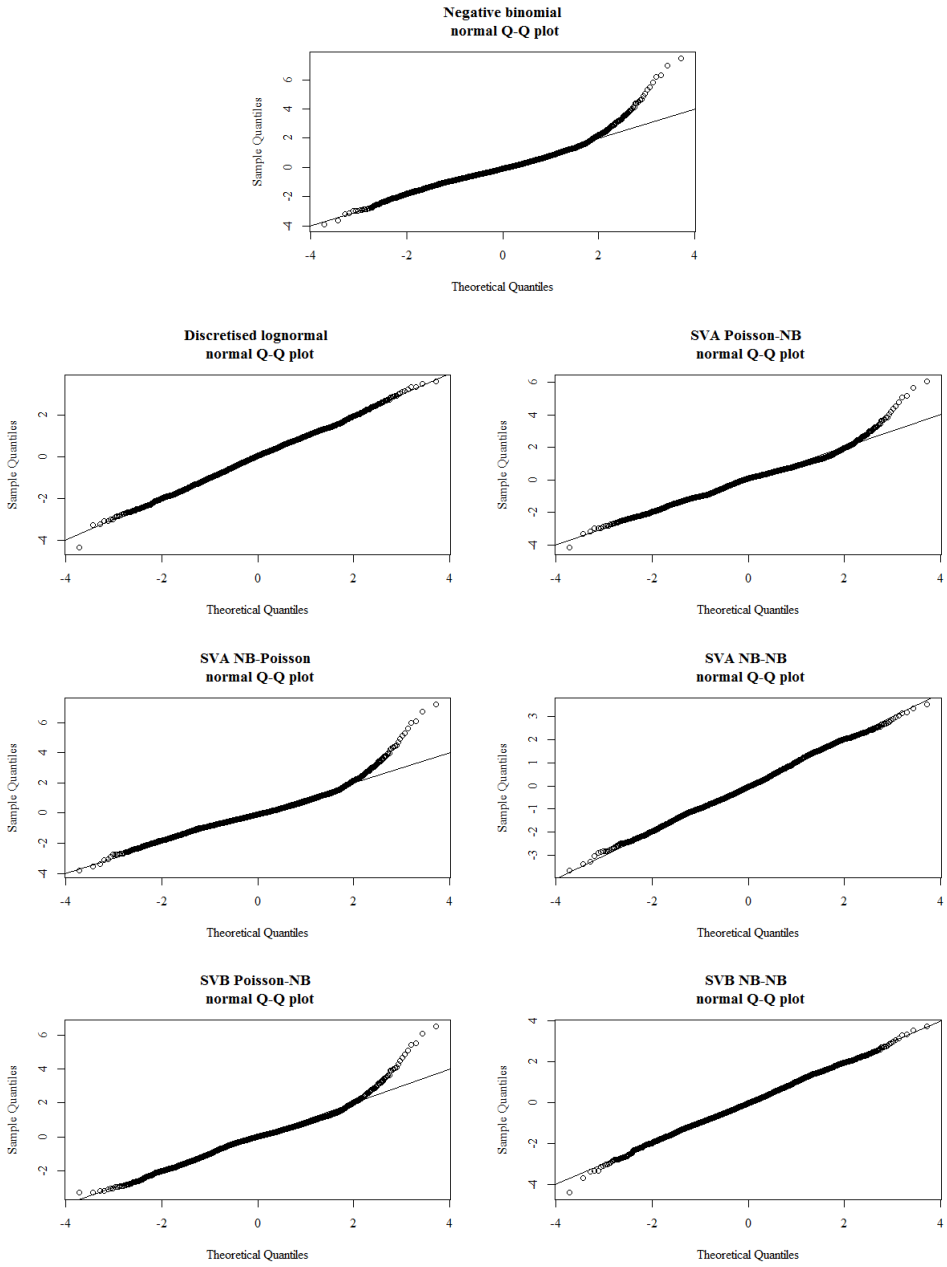

Figure E.14: *Randomised quantile residual plots of models for StatsProb.*

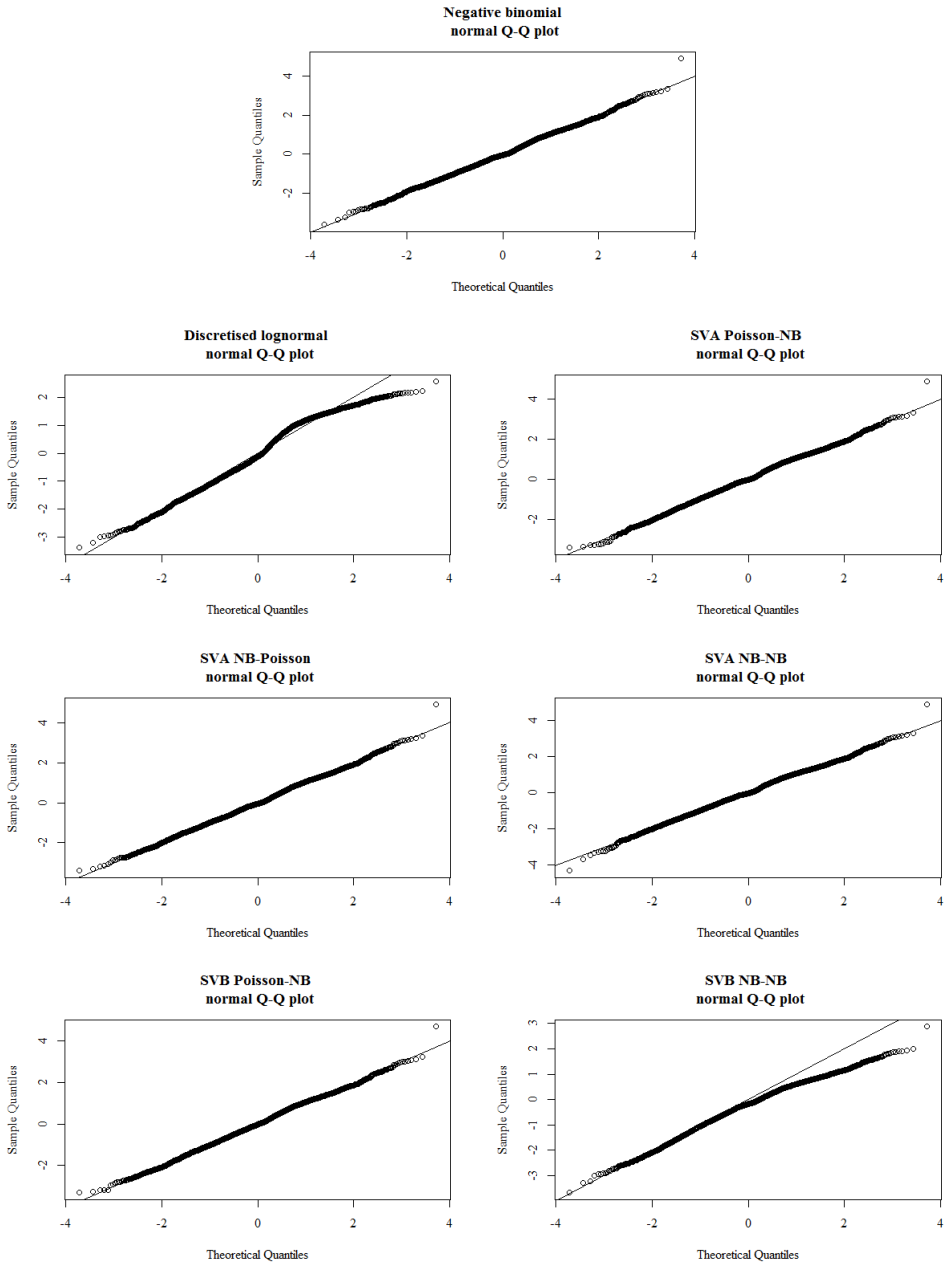

Figure E.15: *Randomised quantile residual plots of models for Rehab.*

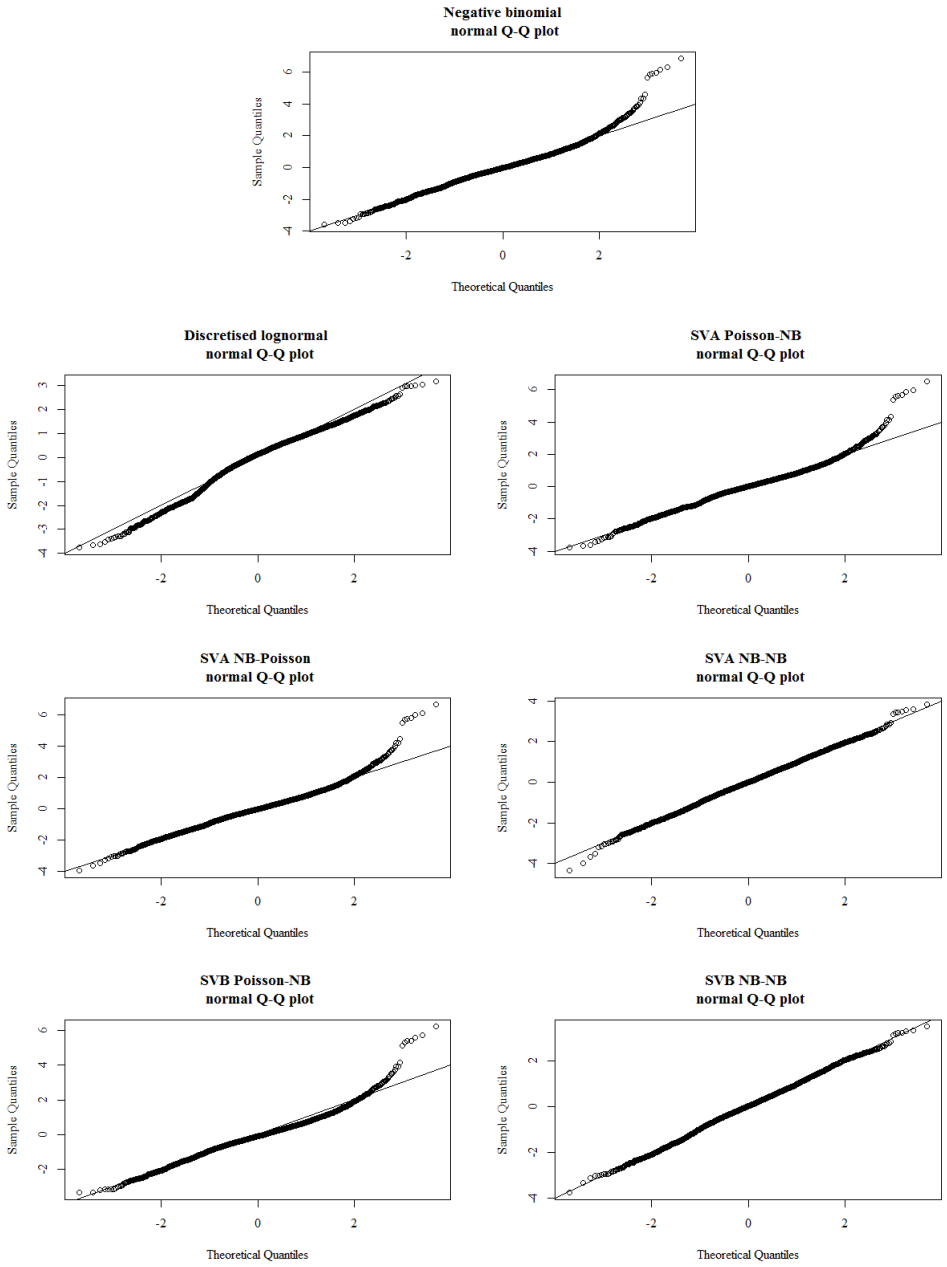

Figure E.16: *Randomised quantile residual plots of models for Oncology.*

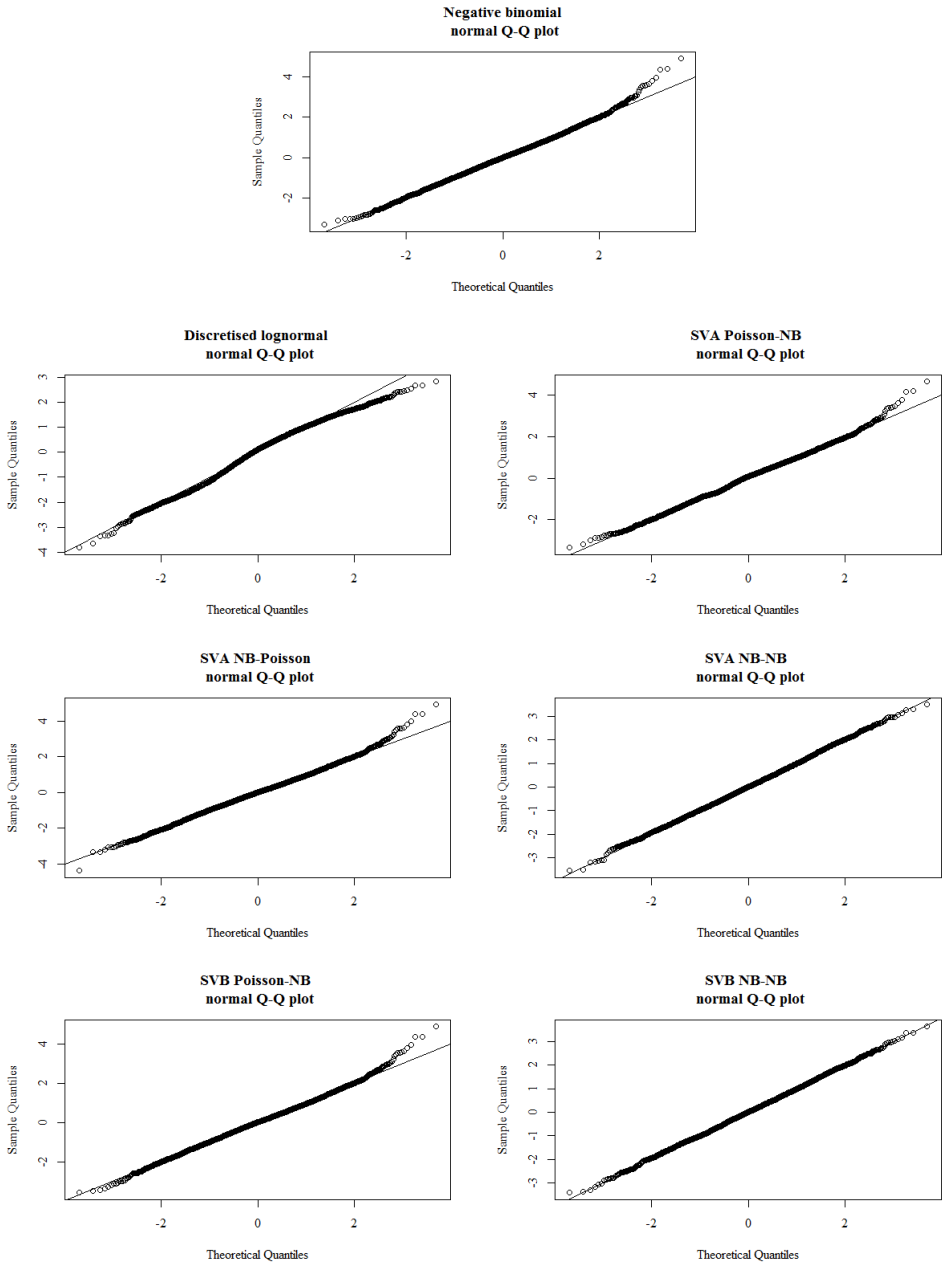

Figure E.17: *Randomised quantile residual plots of models for Logic.*

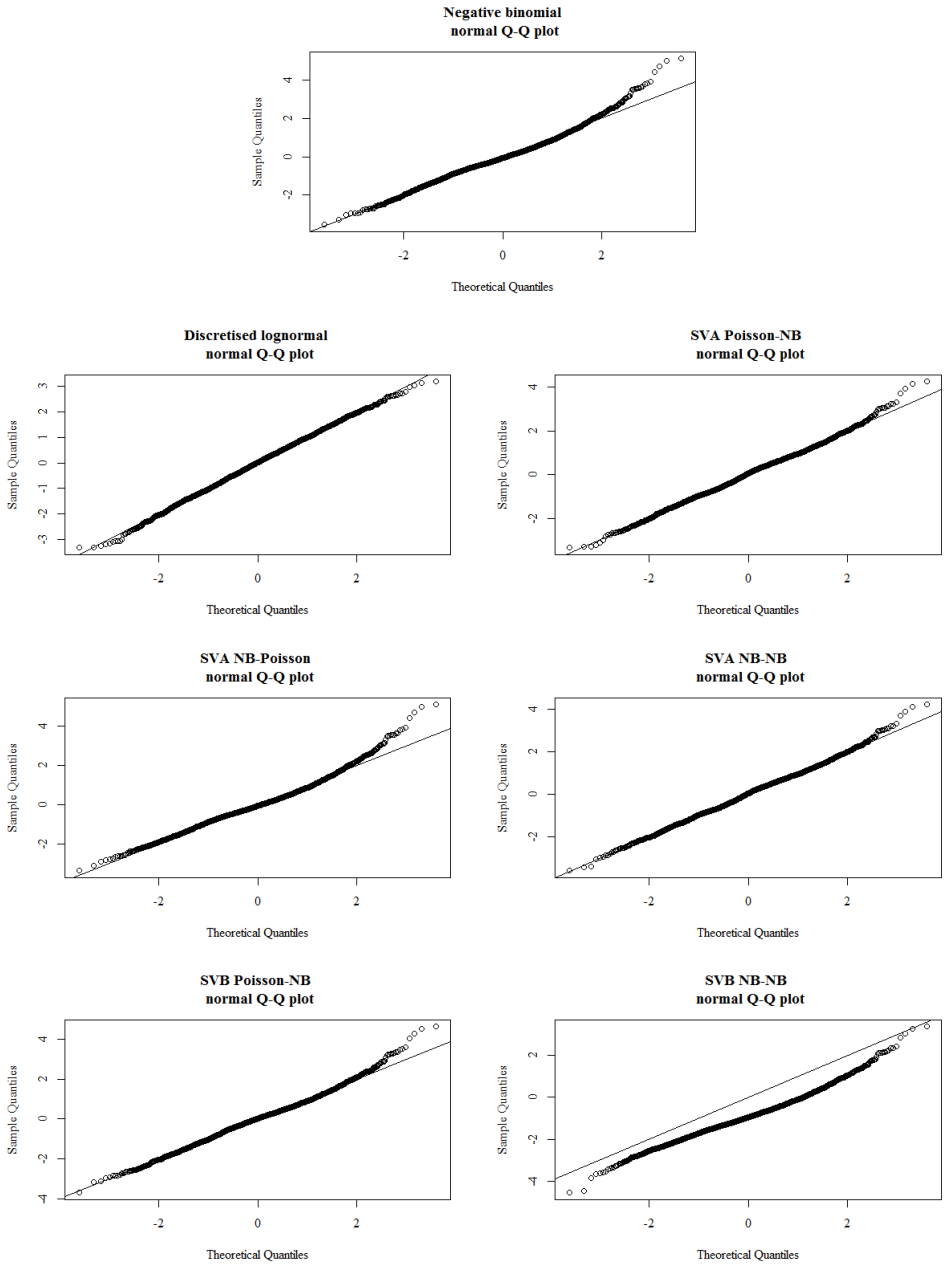

Figure E.18: *Randomised quantile residual plots of models for Dermatology.*

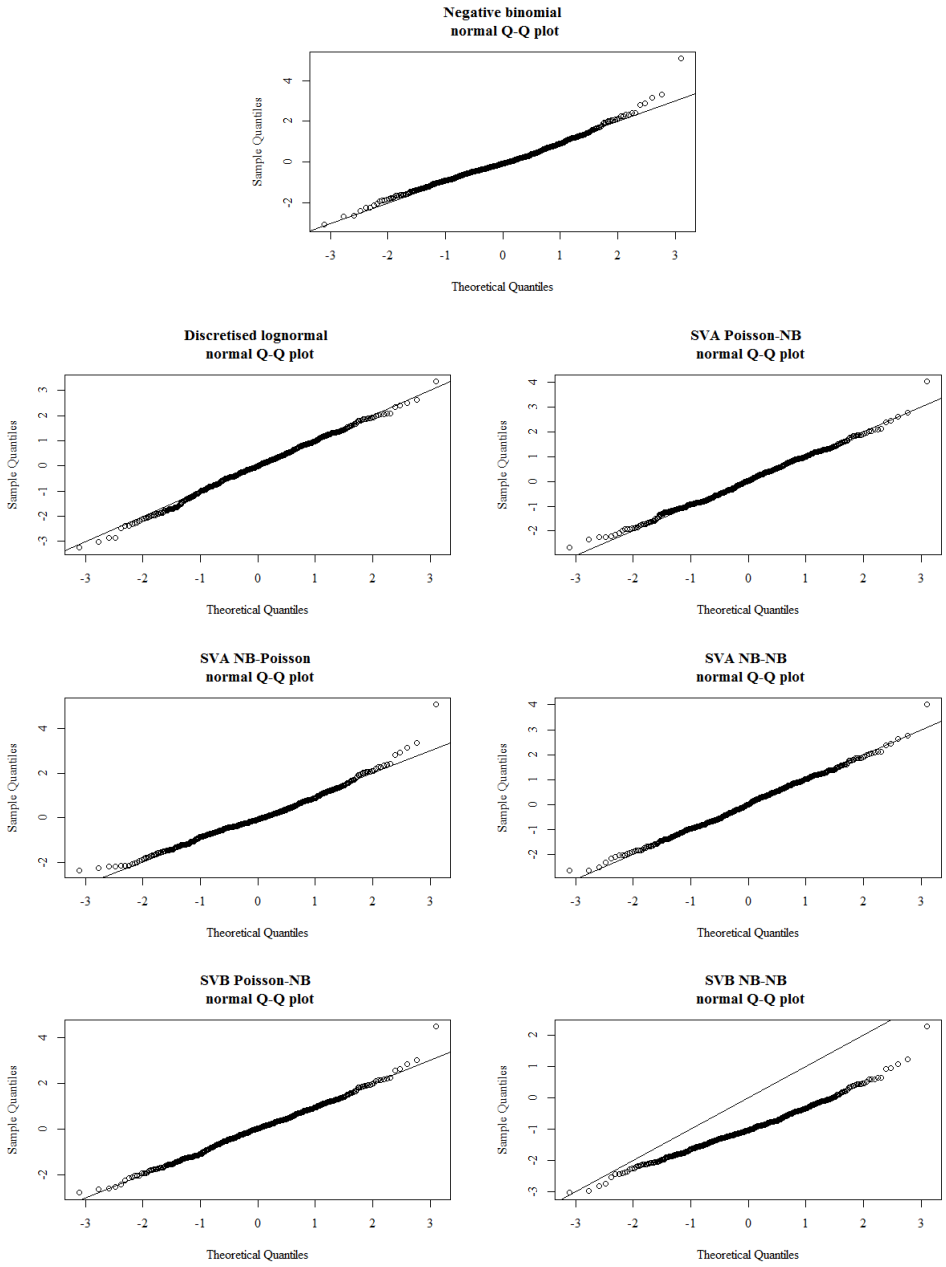

Figure E.19: *Randomised quantile residual plots of models for Algebra.*

Appendix F

Randomised quantile residual plots for citation analysis with covariates

The randomised quantile residual plots for the negative binomial and proposed variant models fitted in Section 5.4 are presented here.

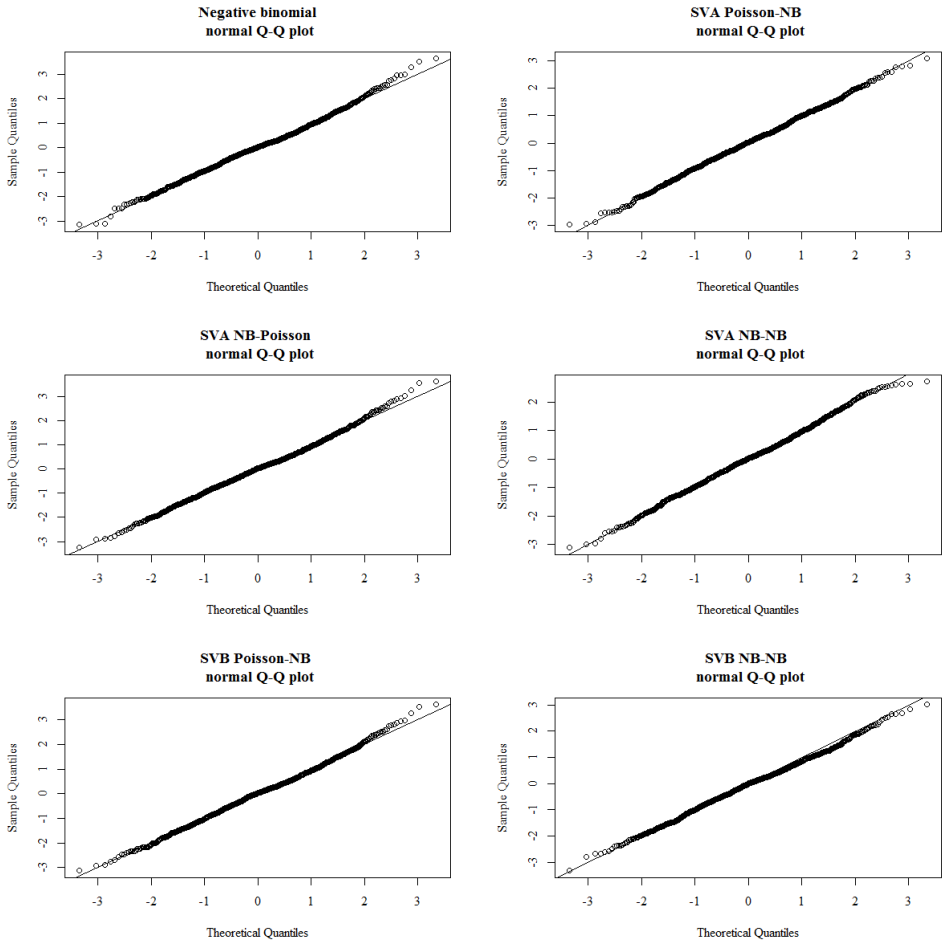

Figure F.1: *Randomised quantile residual plots of models for Archeology.*

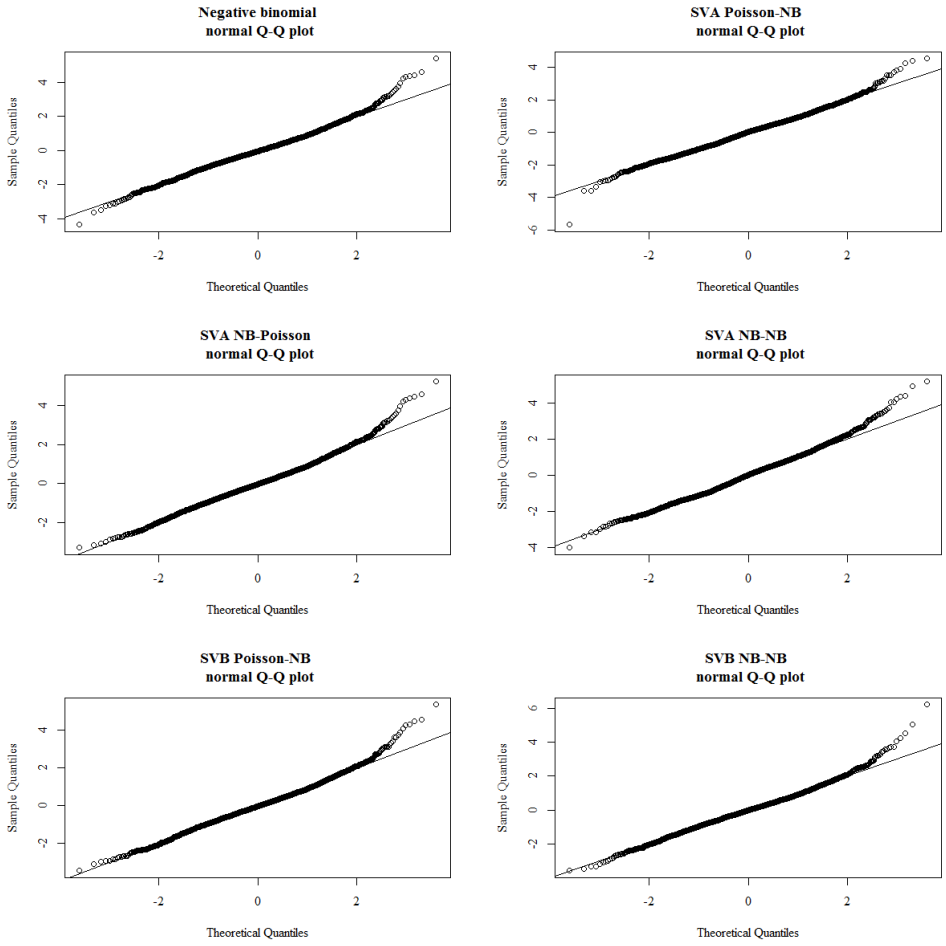

Figure F.2: *Randomised quantile residual plots of models for Biochemistry.*

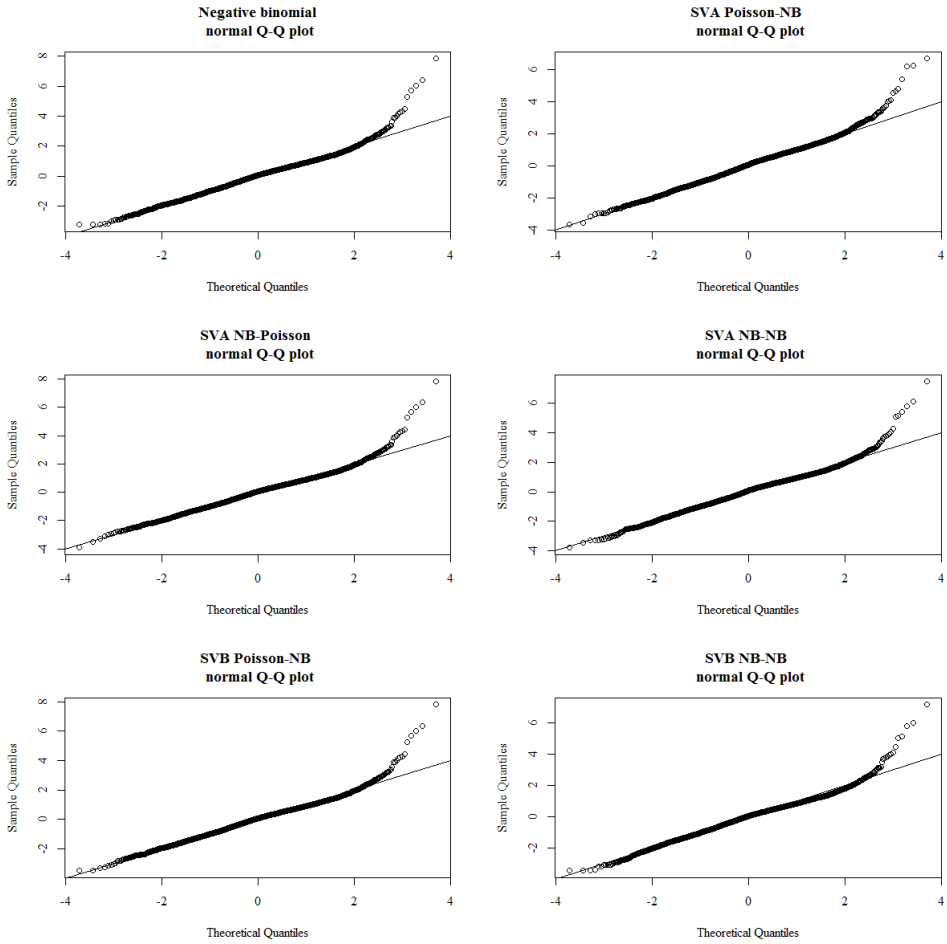

Figure F.3: *Randomised quantile residual plots of models for Biomedical Engineering.*

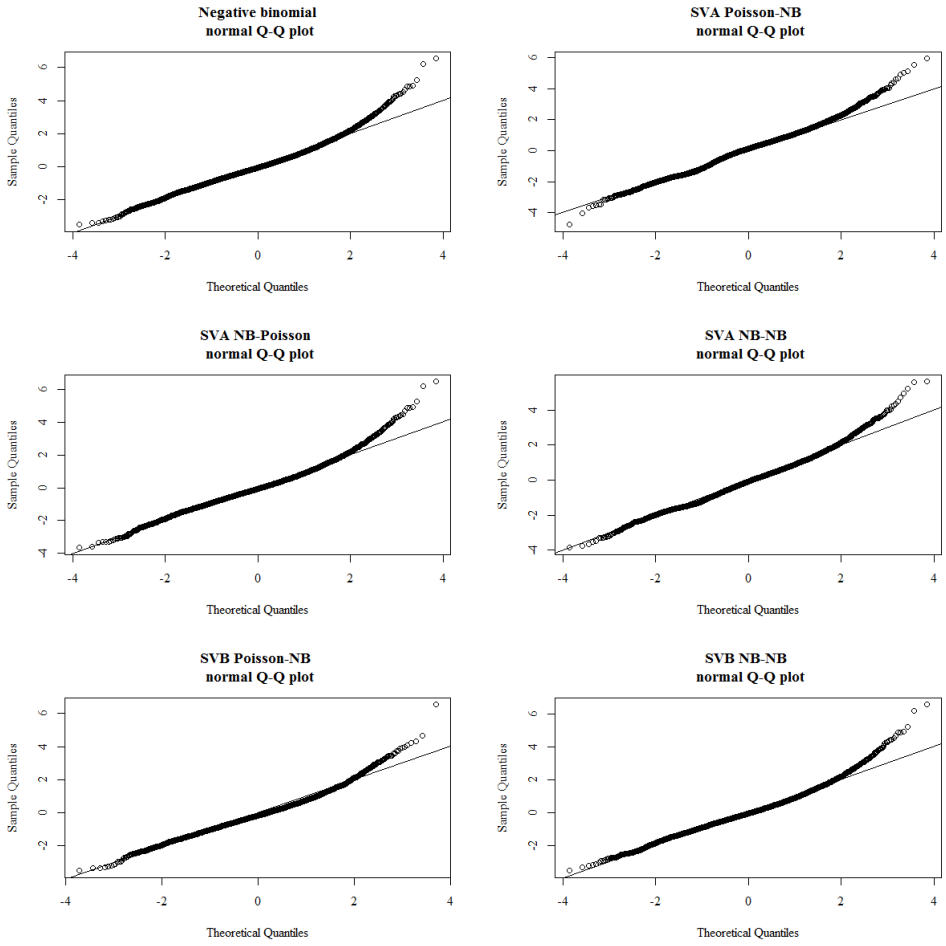

Figure F.4: *Randomised quantile residual plots of models for Biophysics.*

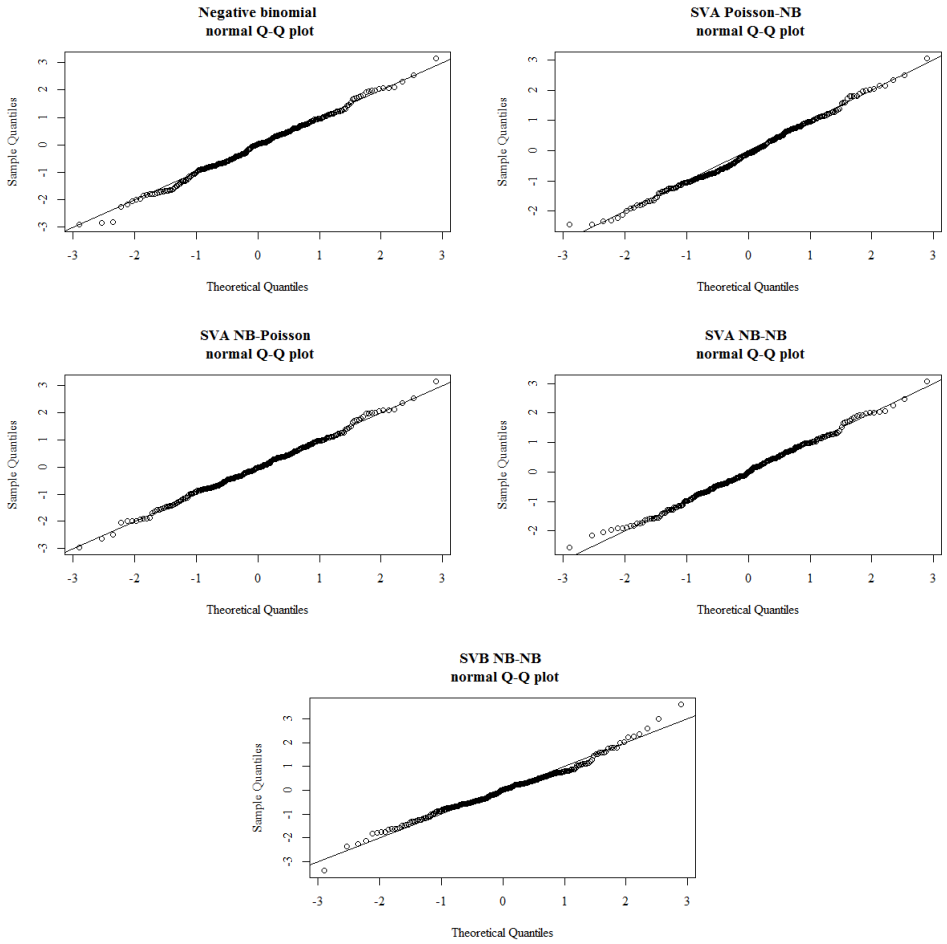

Figure F.5: *Randomised quantile residual plots of models for Care Planning.*

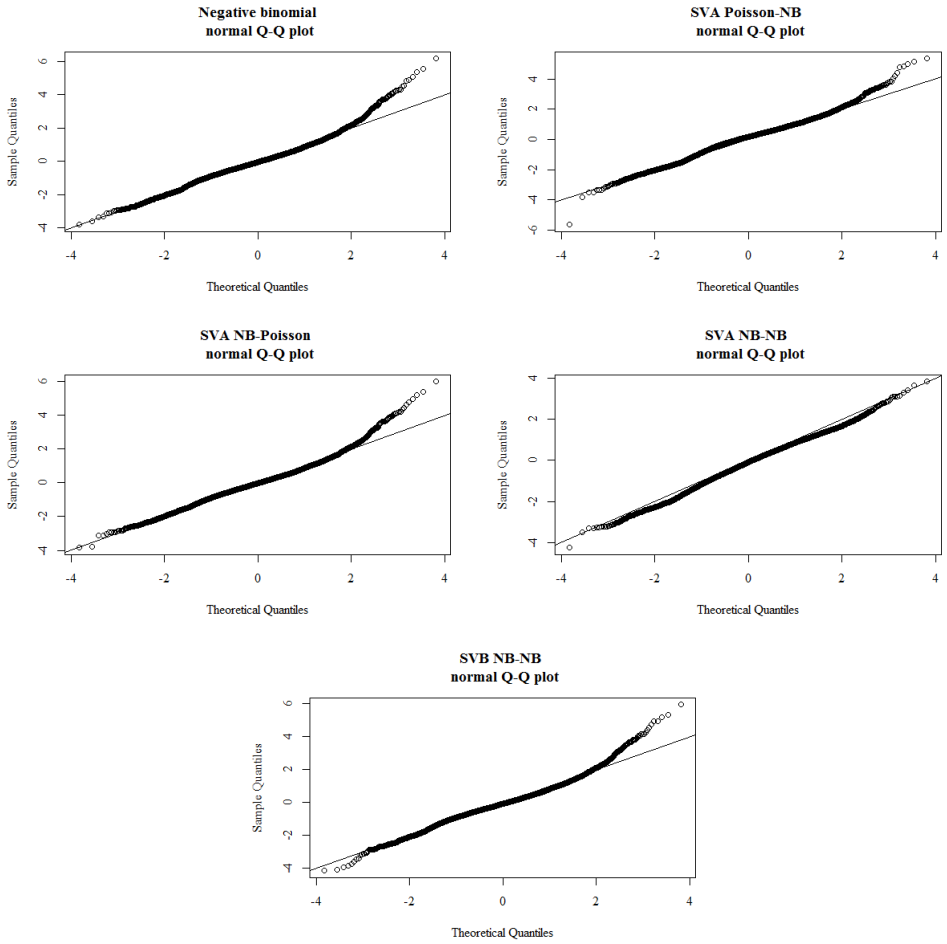

Figure F.6: *Randomised quantile residual plots of models for Cellular and Molecular Neuroscience.*

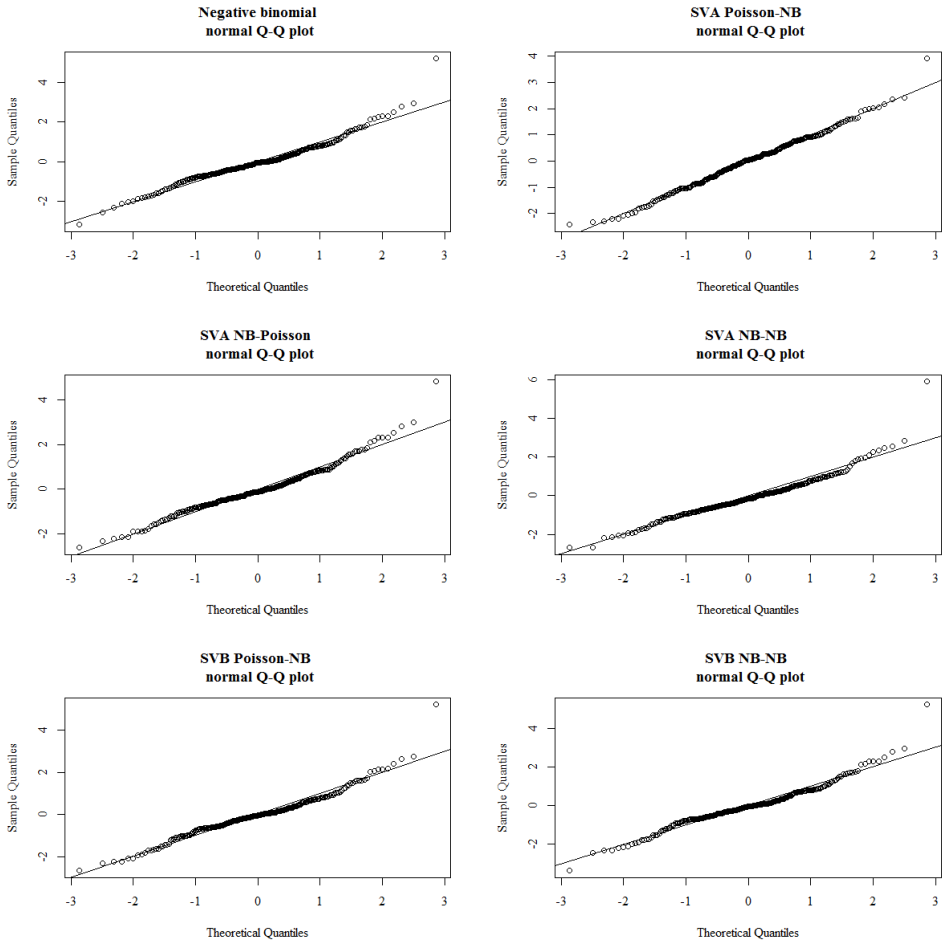

Figure F.7: *Randomised quantile residual plots of models for Chemical Health and Safety.*

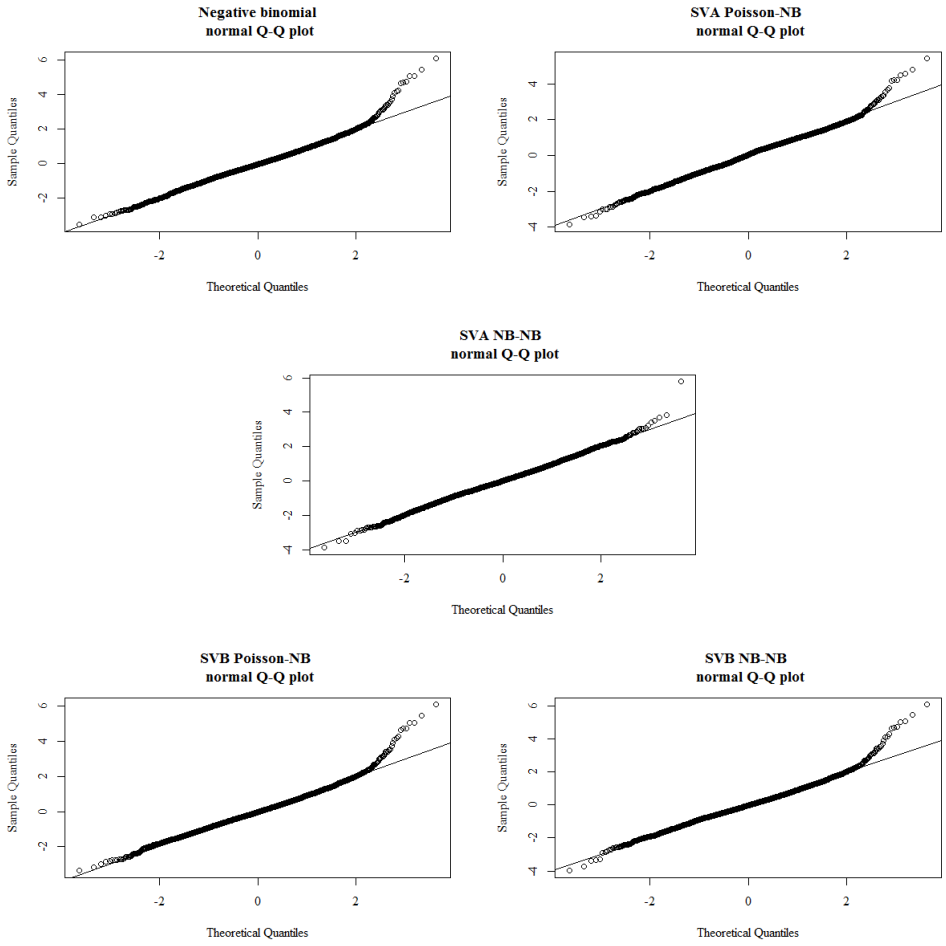

Figure F.8: *Randomised quantile residual plots of models for Computer Graphics and Computer Aided Design.*

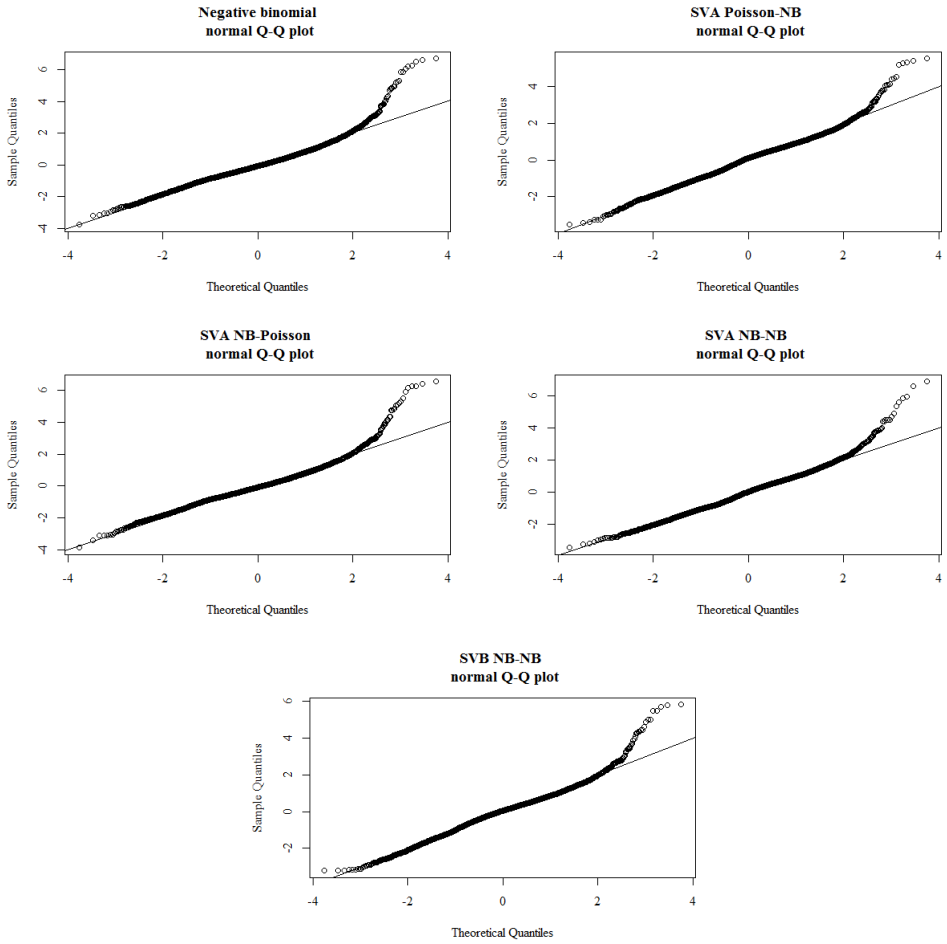

Figure F.9: *Randomised quantile residual plots of models for Condensed Matter Physics.*

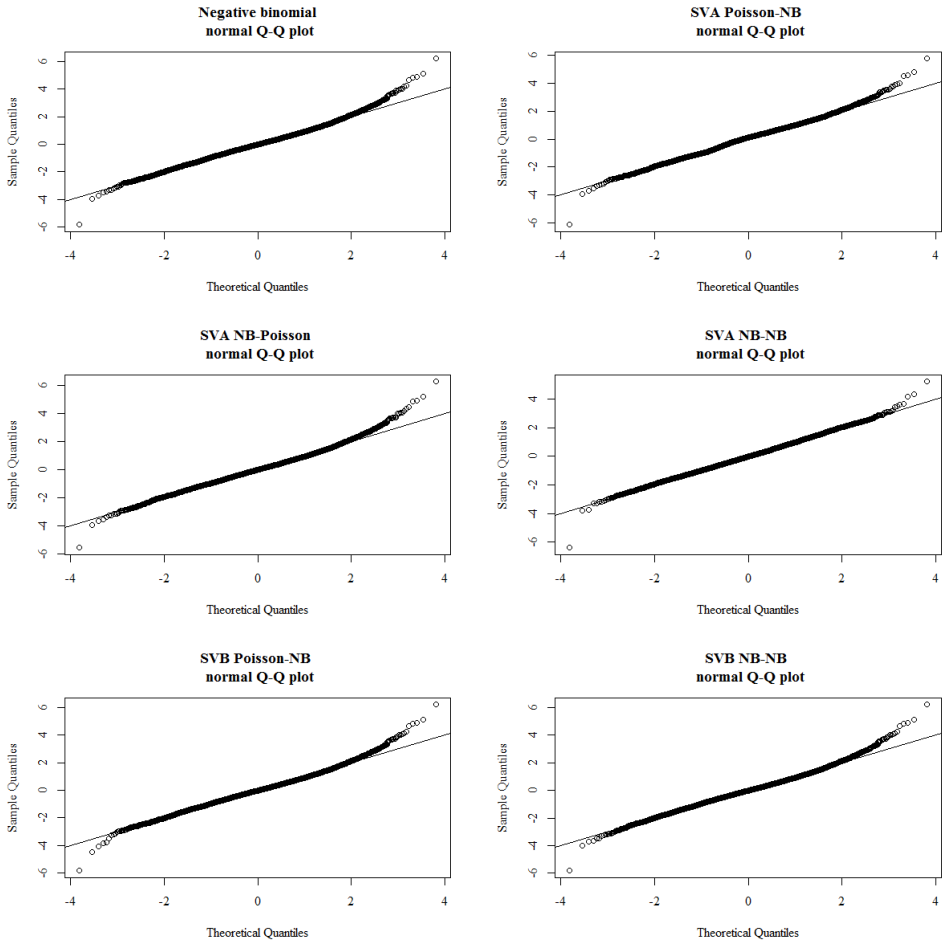

Figure F.10: *Randomised quantile residual plots of models for Developmental and Educational Psychology.*

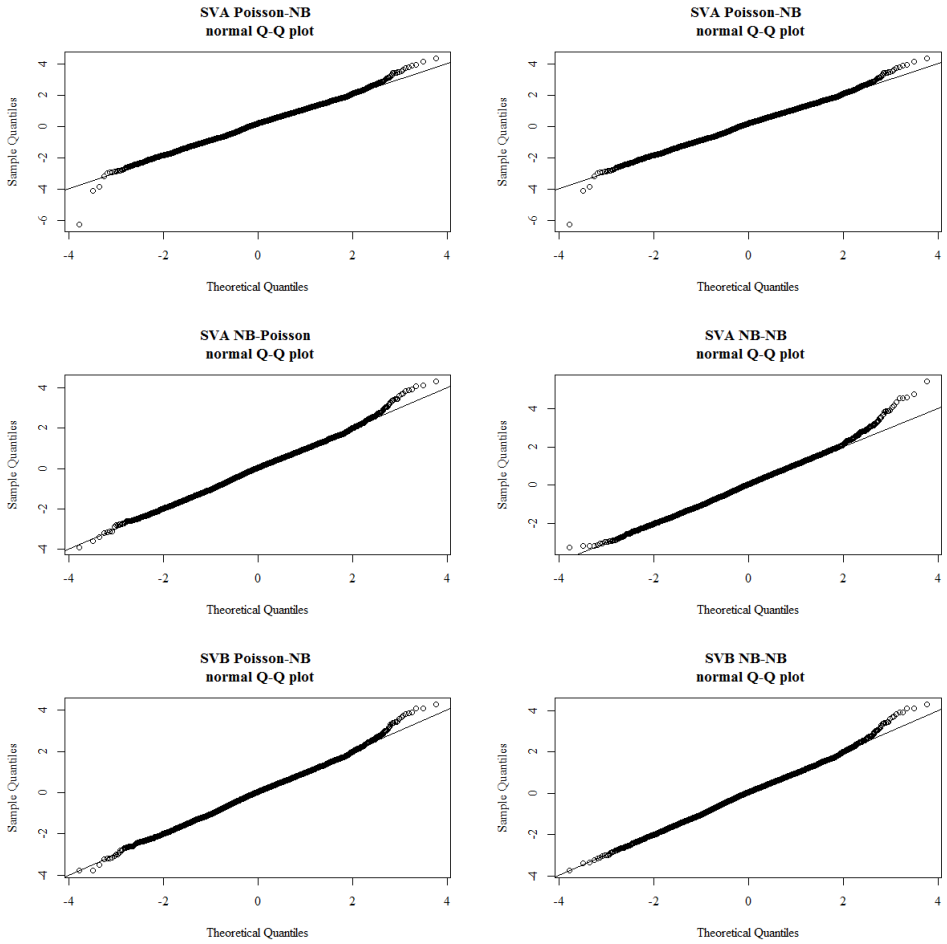

Figure F.11: *Randomised quantile residual plots of models for Earth Surface Processes.*

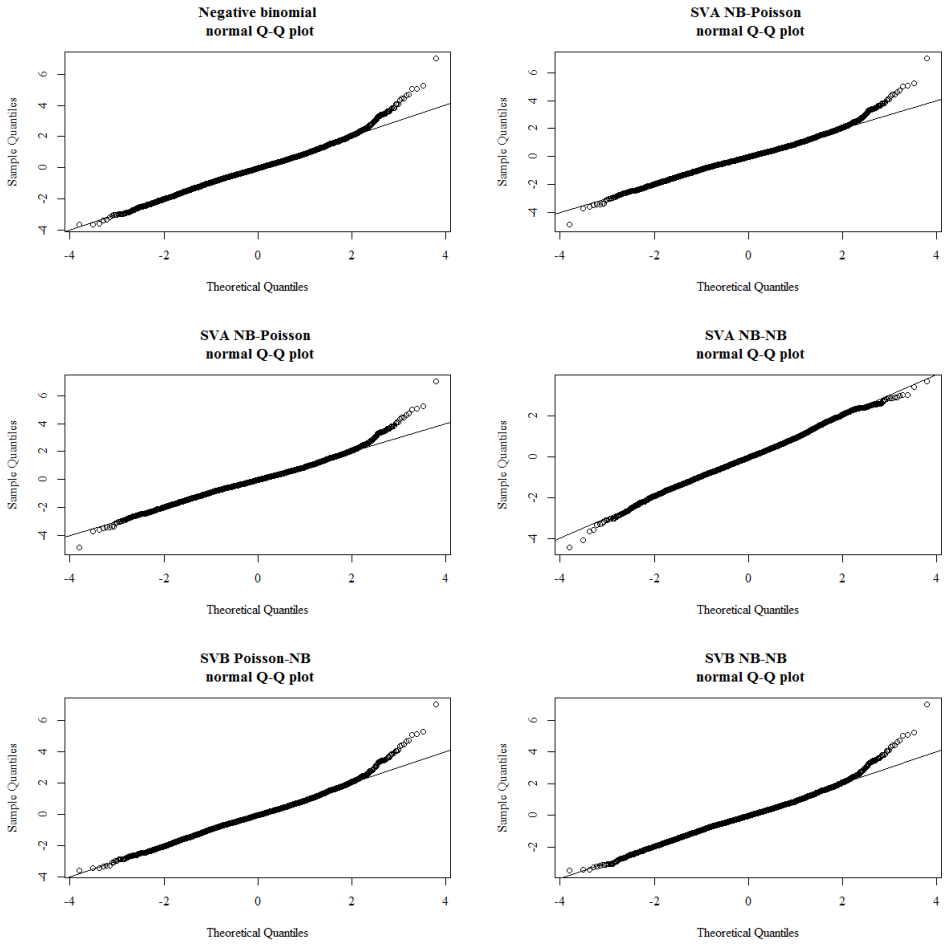

Figure F.12: *Randomised quantile residual plots of models for Education.*

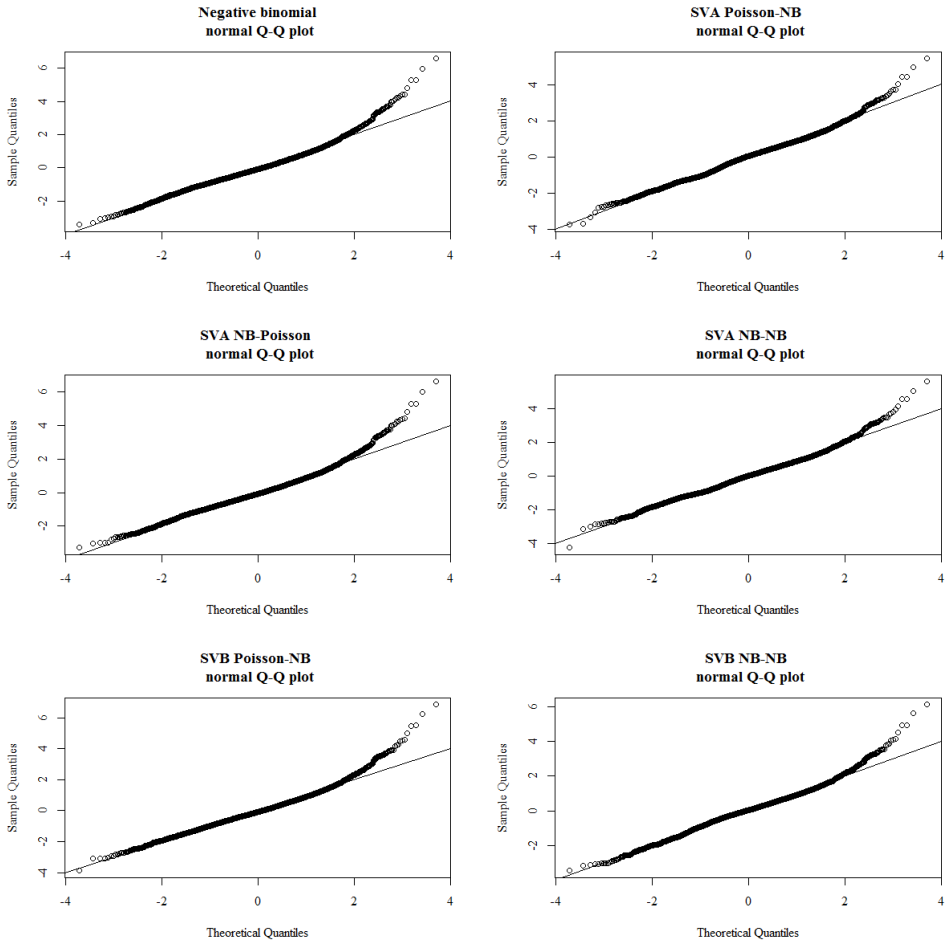

Figure F.13: *Randomised quantile residual plots of models for Electronic Optical and Magnetic Materials.*

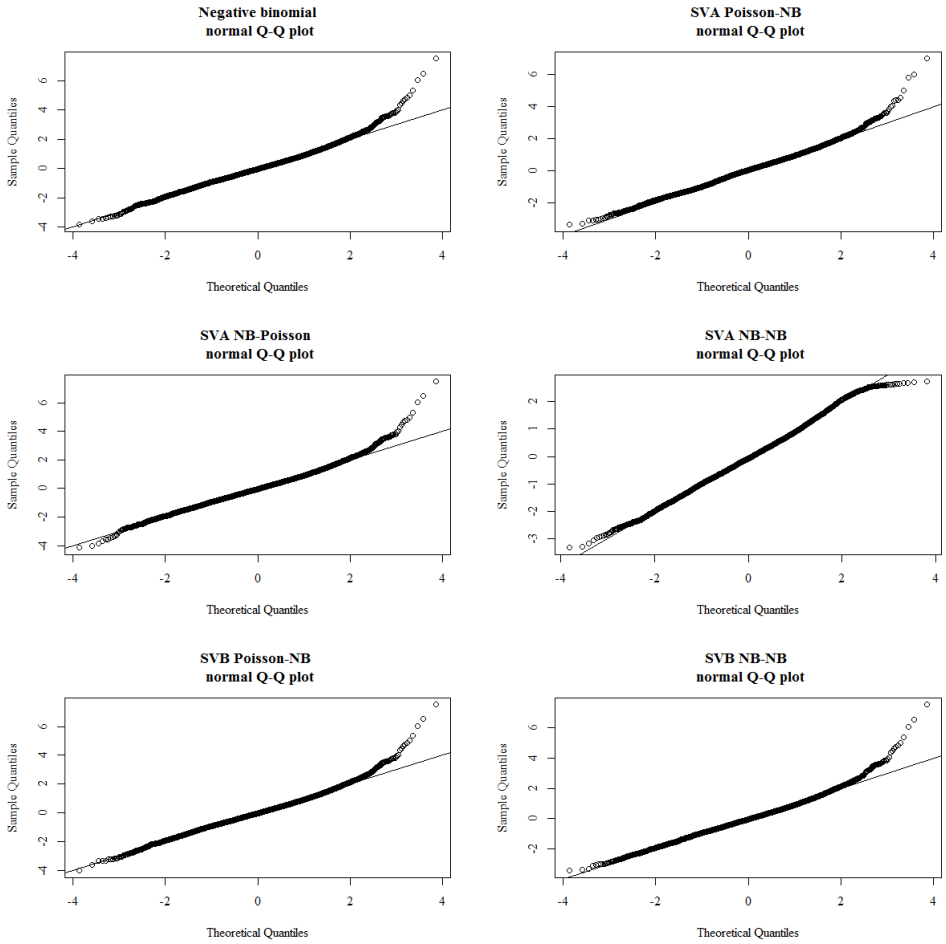

Figure F.14: *Randomised quantile residual plots of models for Environmental Chemistry.*

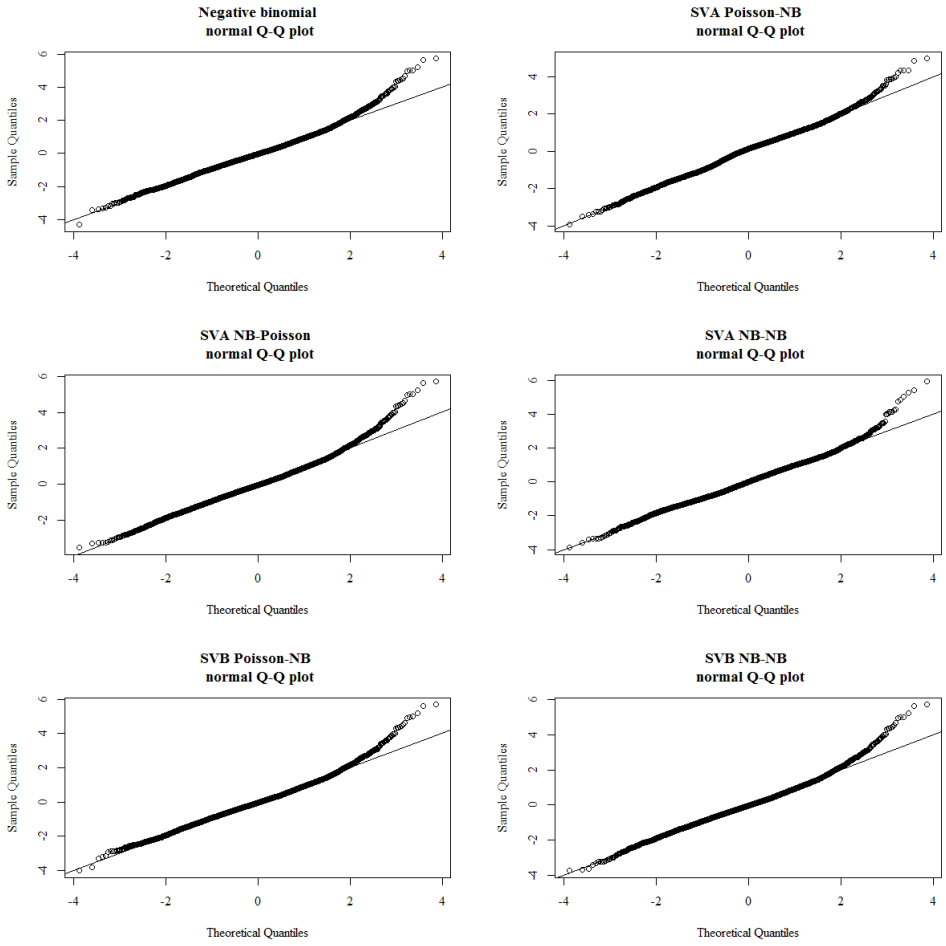

Figure F.15: *Randomised quantile residual plots of models for Inorganic Chemistry.*

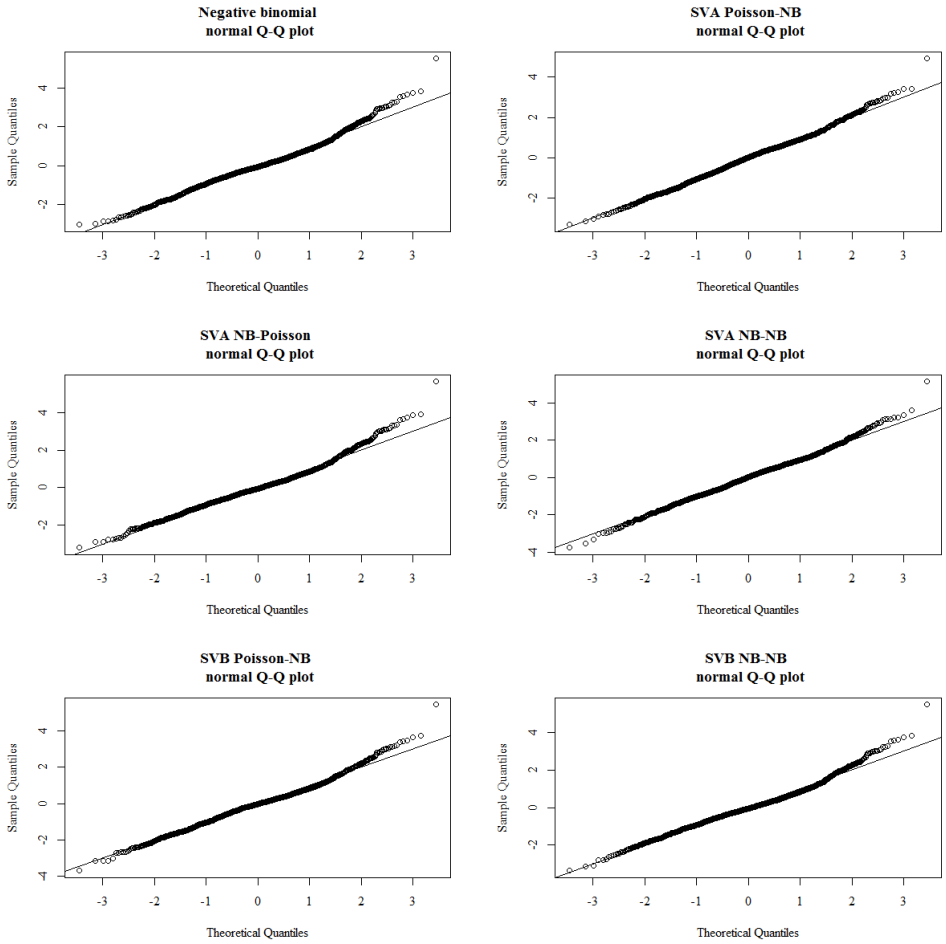

Figure F.16: *Randomised quantile residual plots of models for Management Information Systems.*

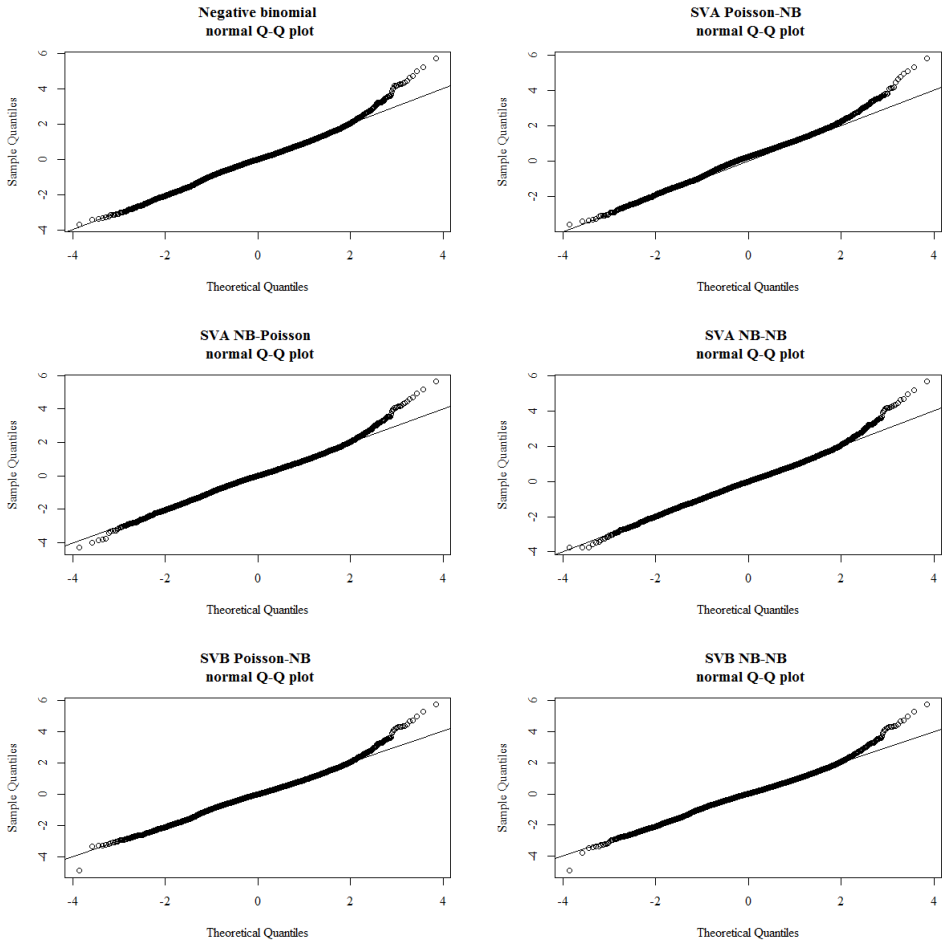

Figure F.17: *Randomised quantile residual plots of models for Microbiology.*

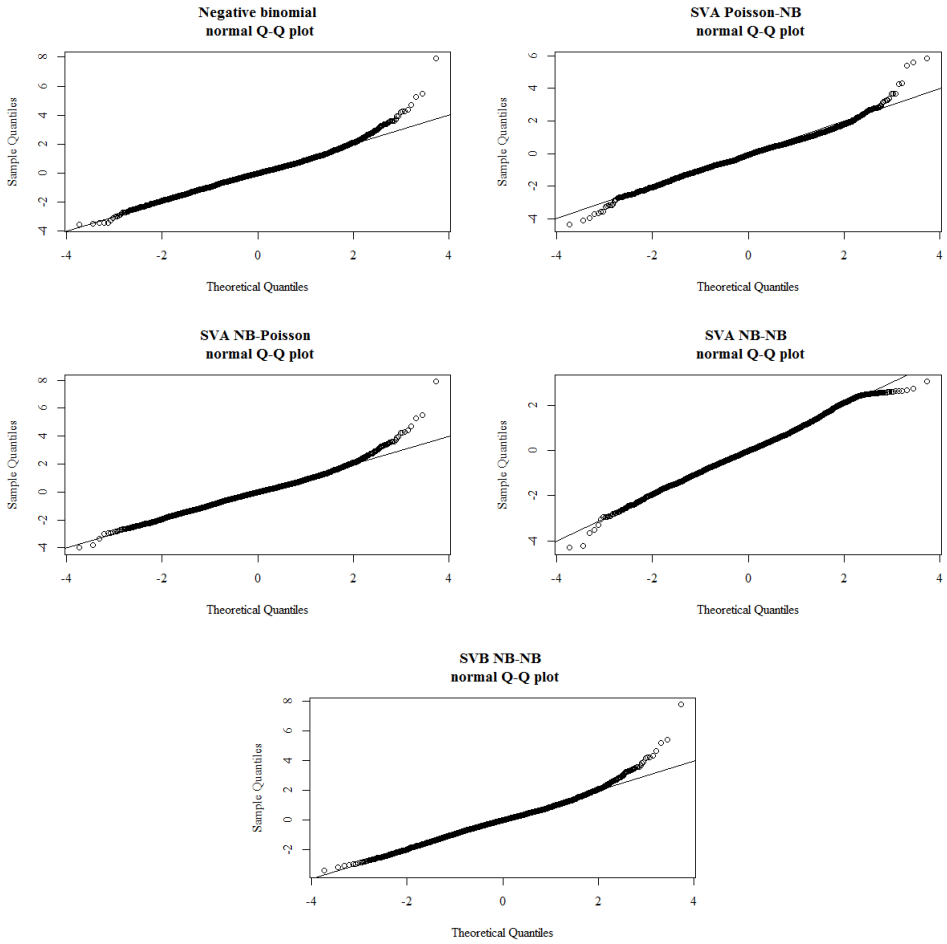

Figure F.18: *Randomised quantile residual plots of models for Nuclear Energy and Engineering.*

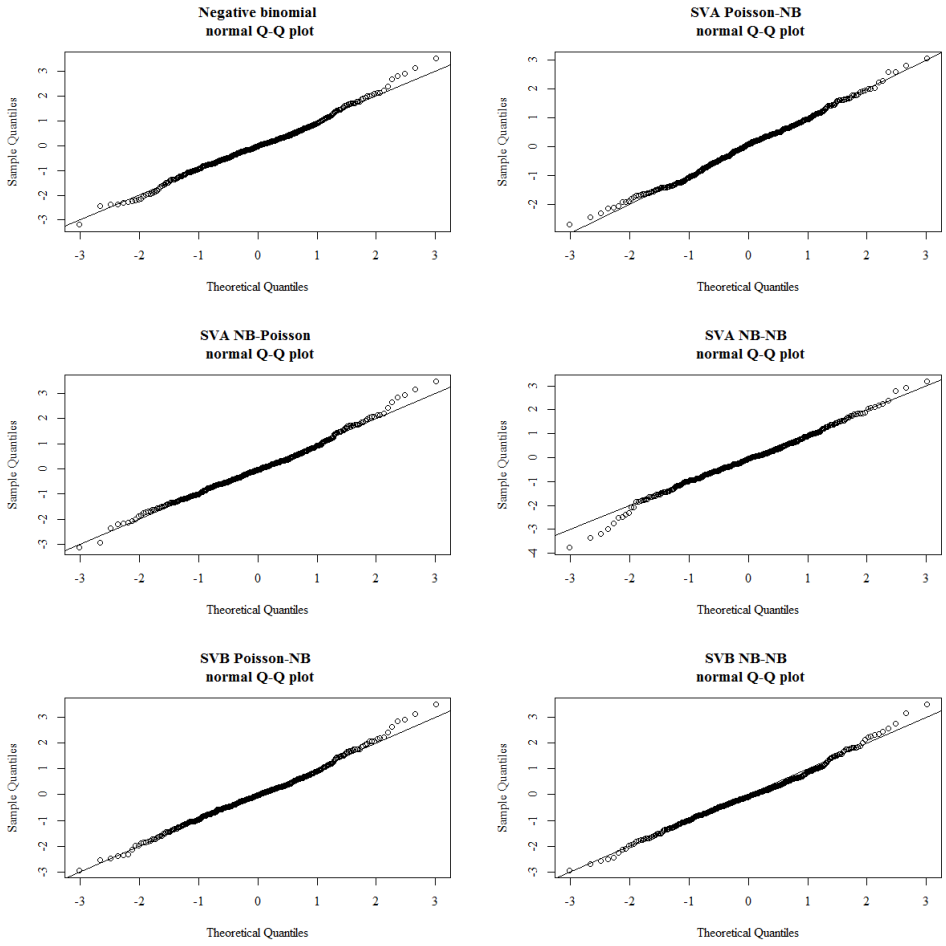

Figure F.19: *Randomised quantile residual plots of models for Oral Surgery.*

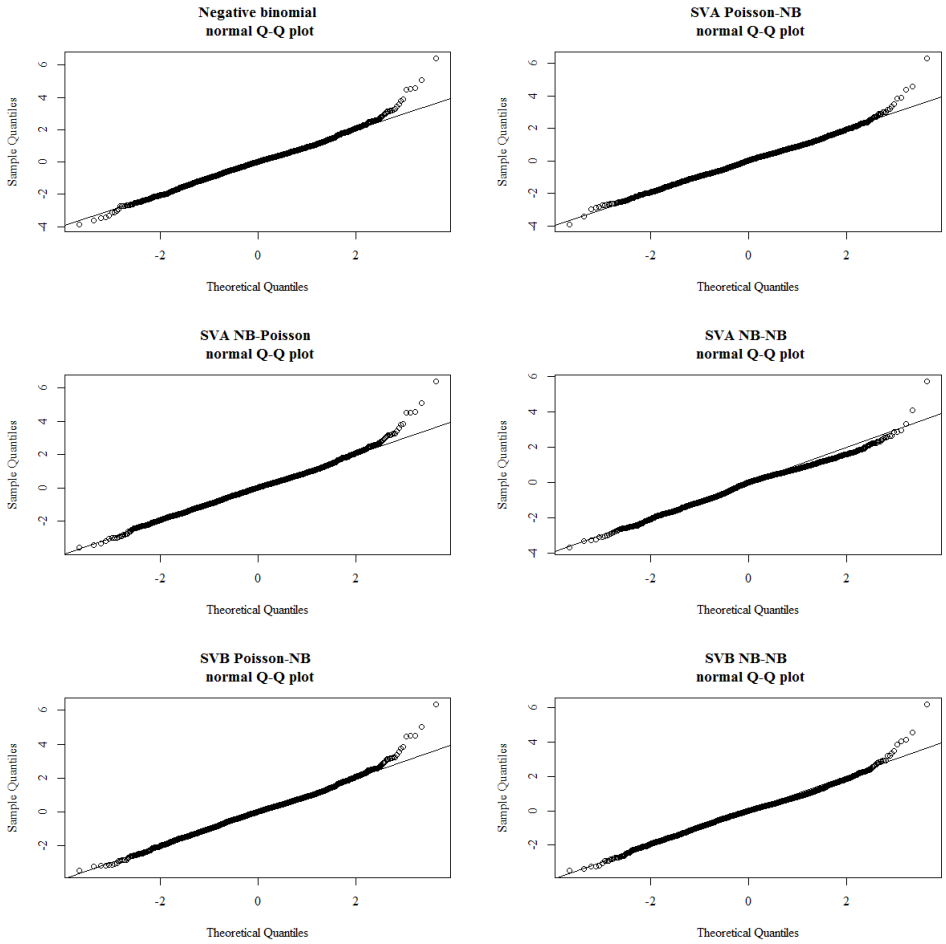

Figure F.20: *Randomised quantile residual plots of models for Pharmacology.*

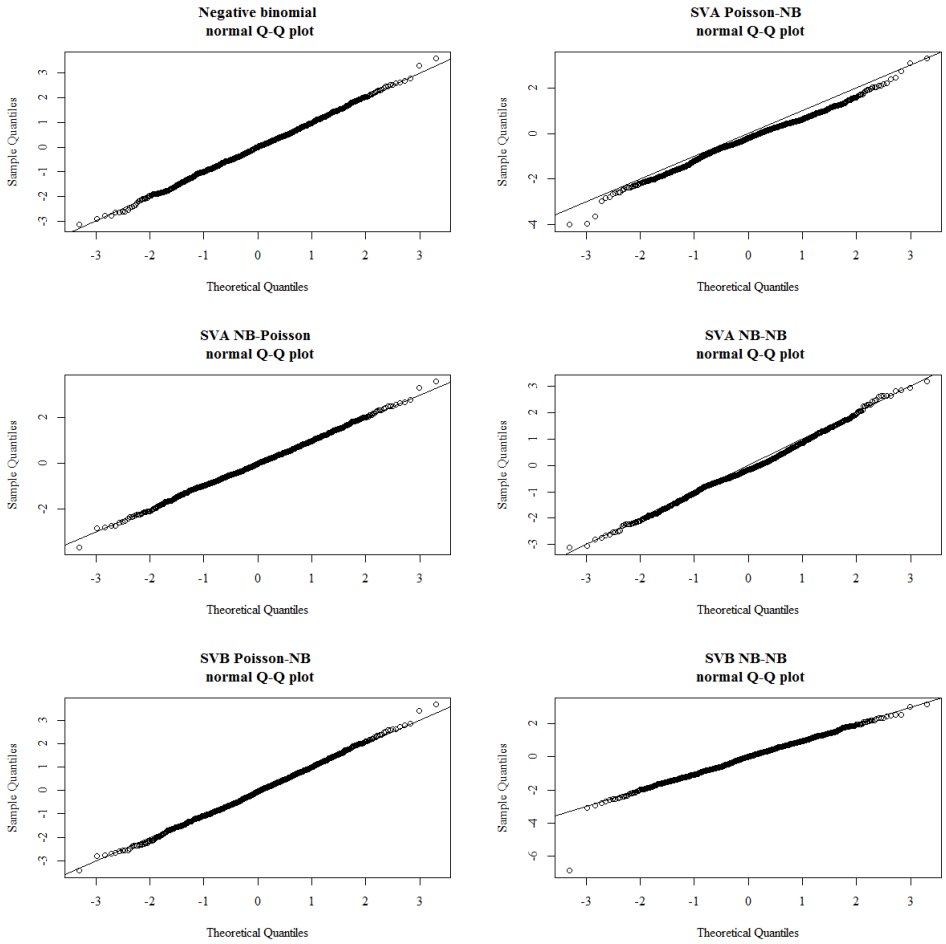

Figure F.21: *Randomised quantile residual plots of models for Small Animals.*

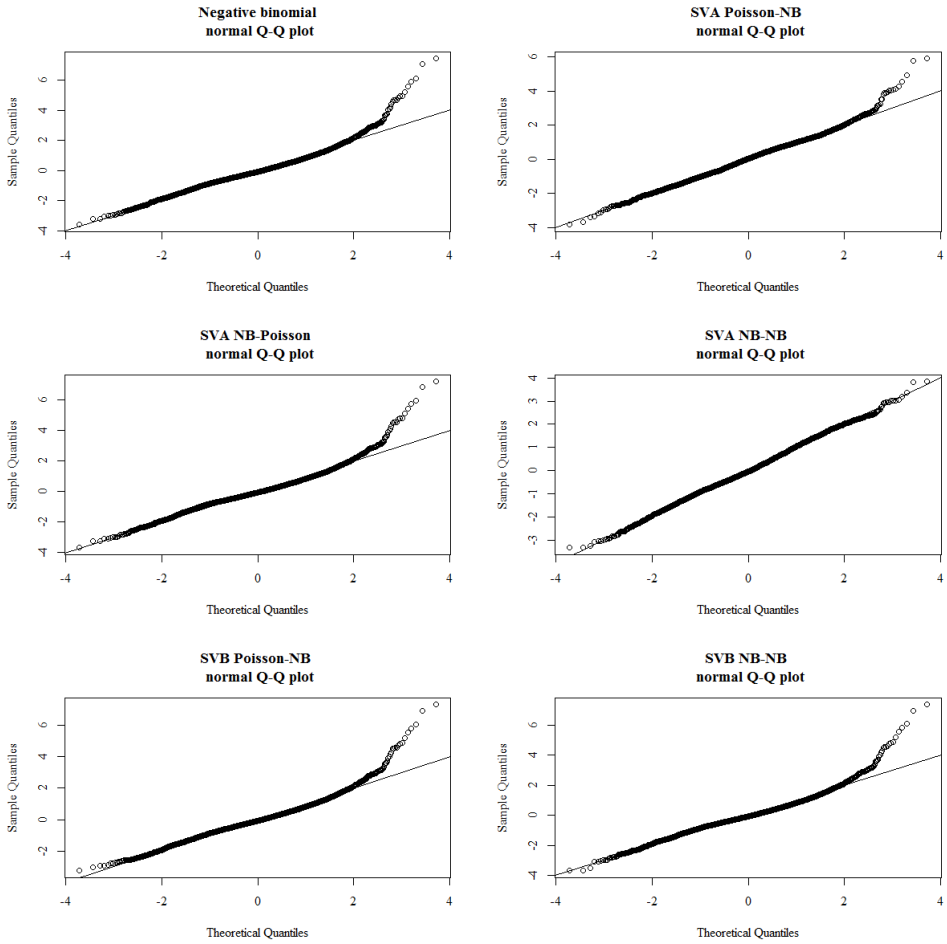

Figure F.22: *Randomised quantile residual plots of models for Statistics Probability and Uncertainty.*

Appendix G

Christmas tree plots for citation analysis with no covariates

This section presents the Christmas tree plots for some models fitted to citation counts in Section 5.3. In all cases, the orange lines denote the 90% fluctuation intervals while the green crosses are the observed median adjusted counts.

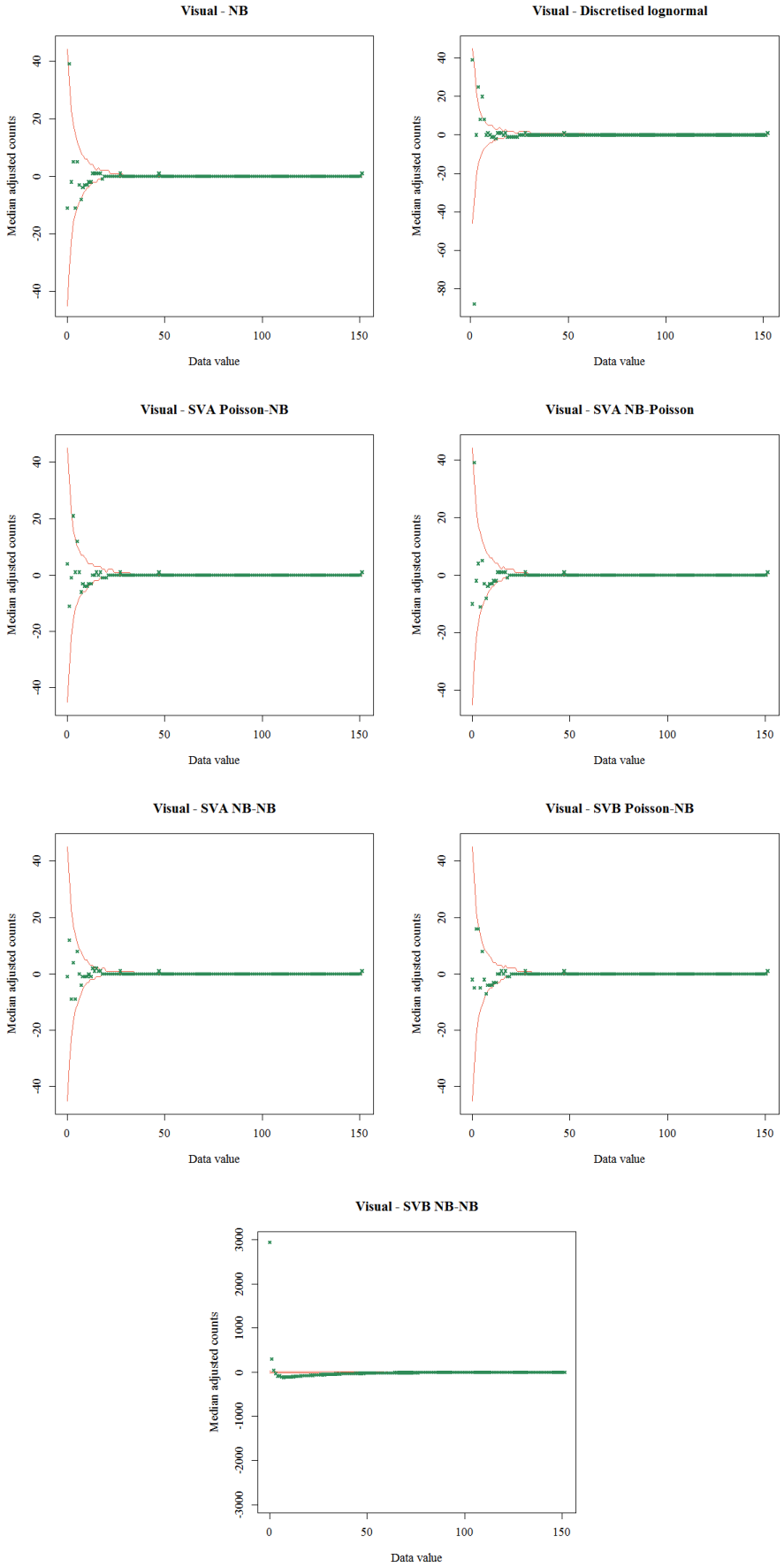

Figure G.1: *Christmas tree plots for Visual.*

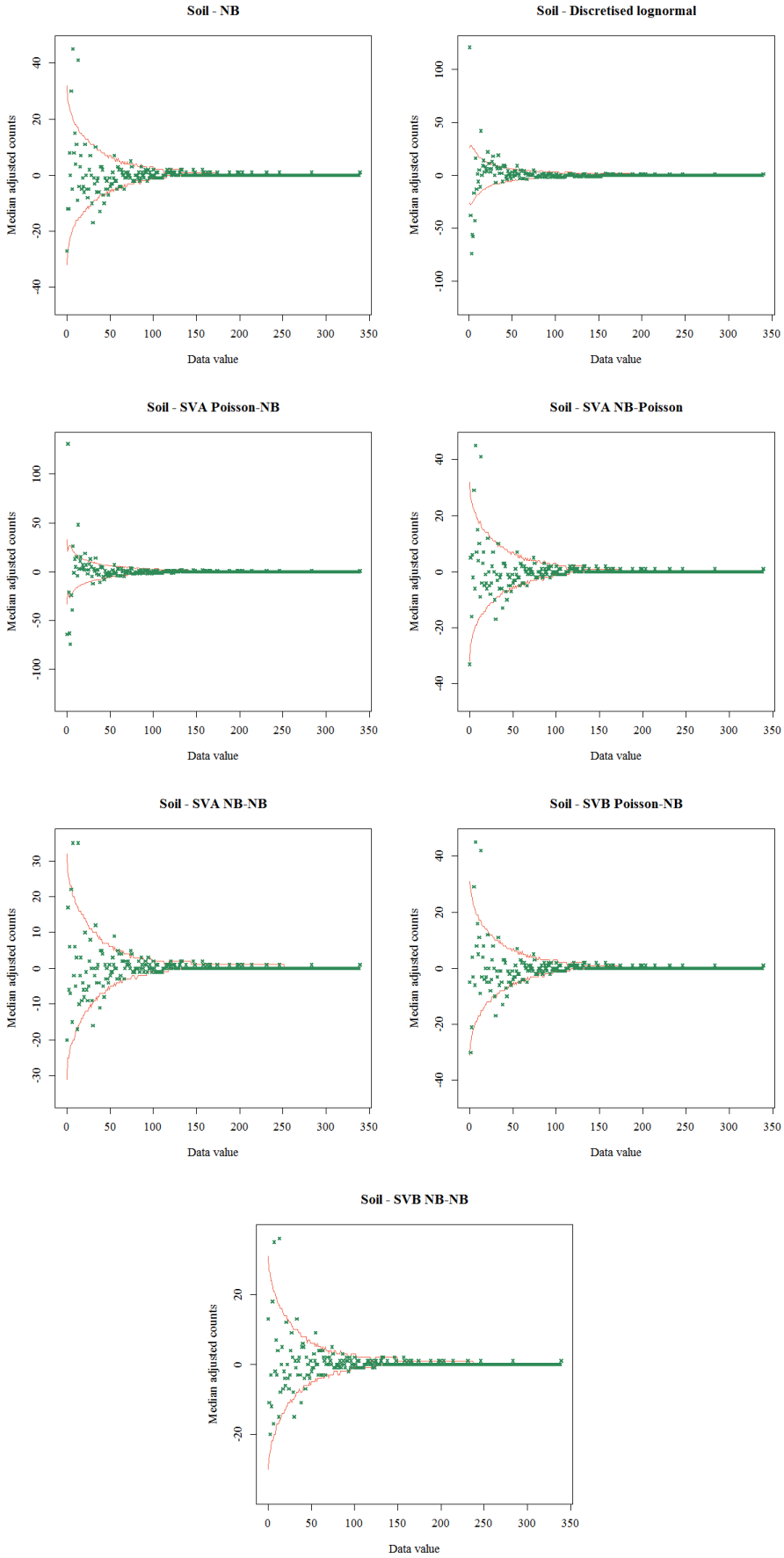

Figure G.2: *Christmas tree plots for Soil.*

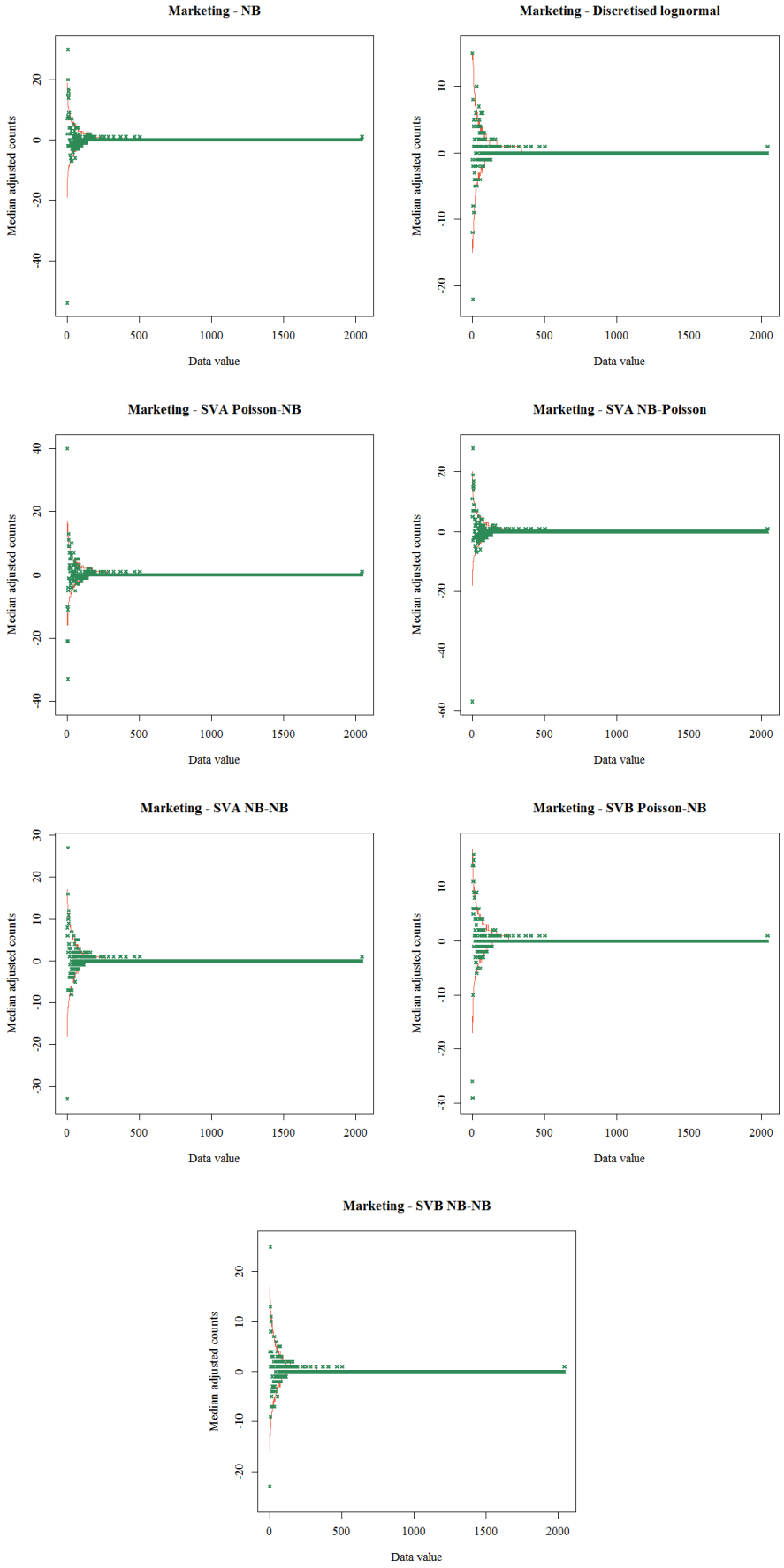

Figure G.3: *Christmas tree plots for Marketing.*

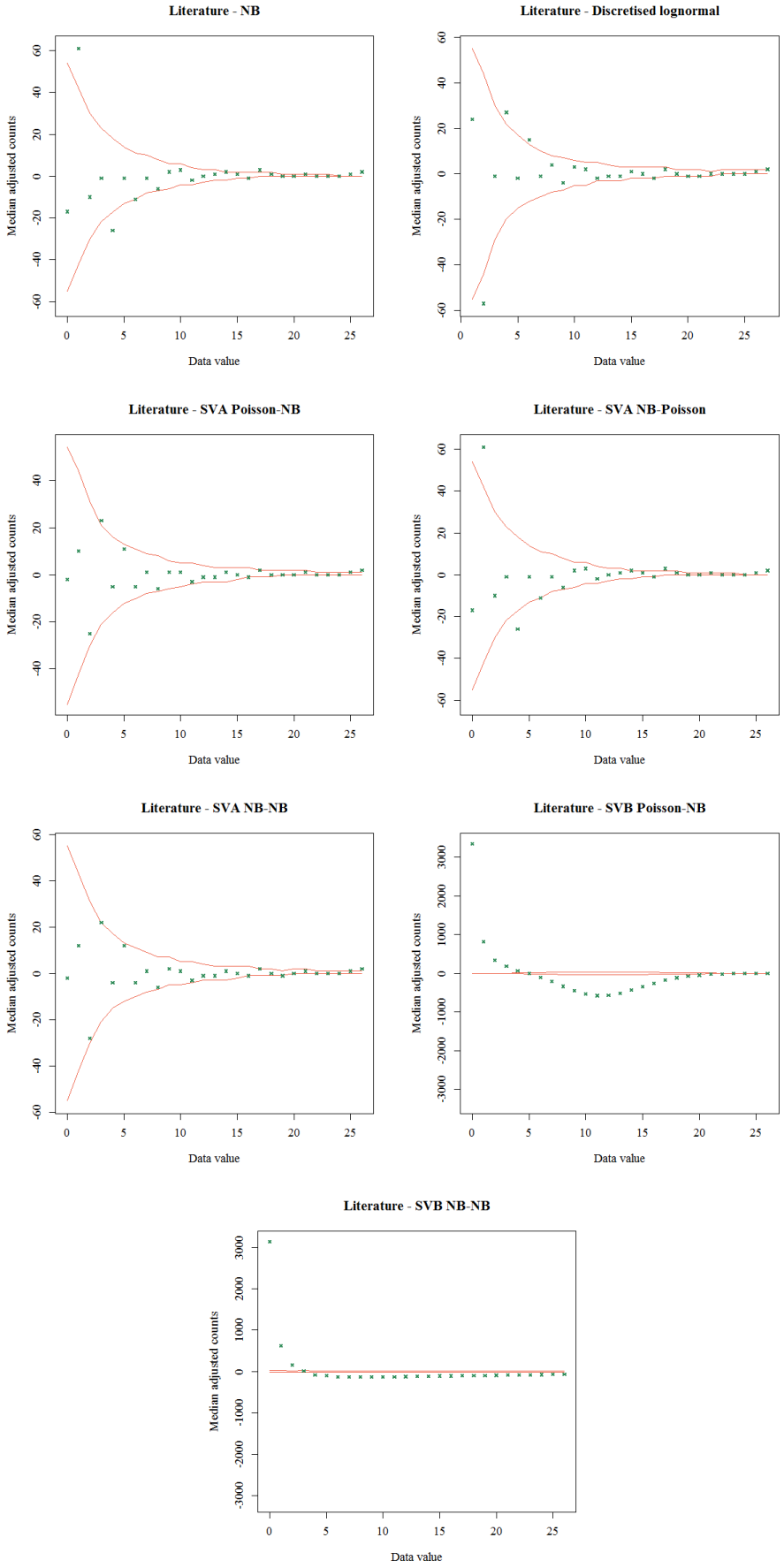

Figure G.4: *Christmas tree plots for Literature.*

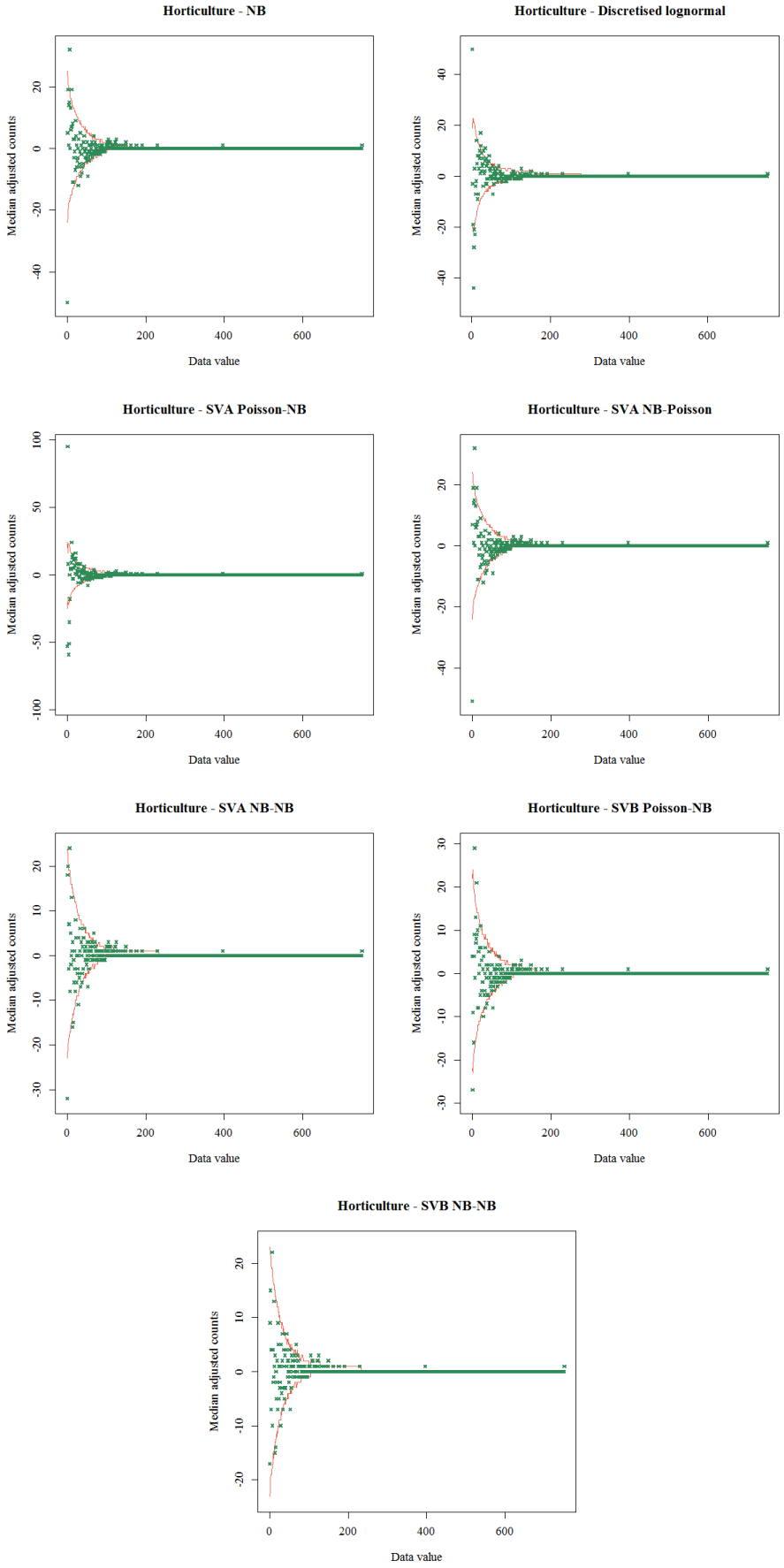

Figure G.5: *Christmas tree plots for Horticulture.*

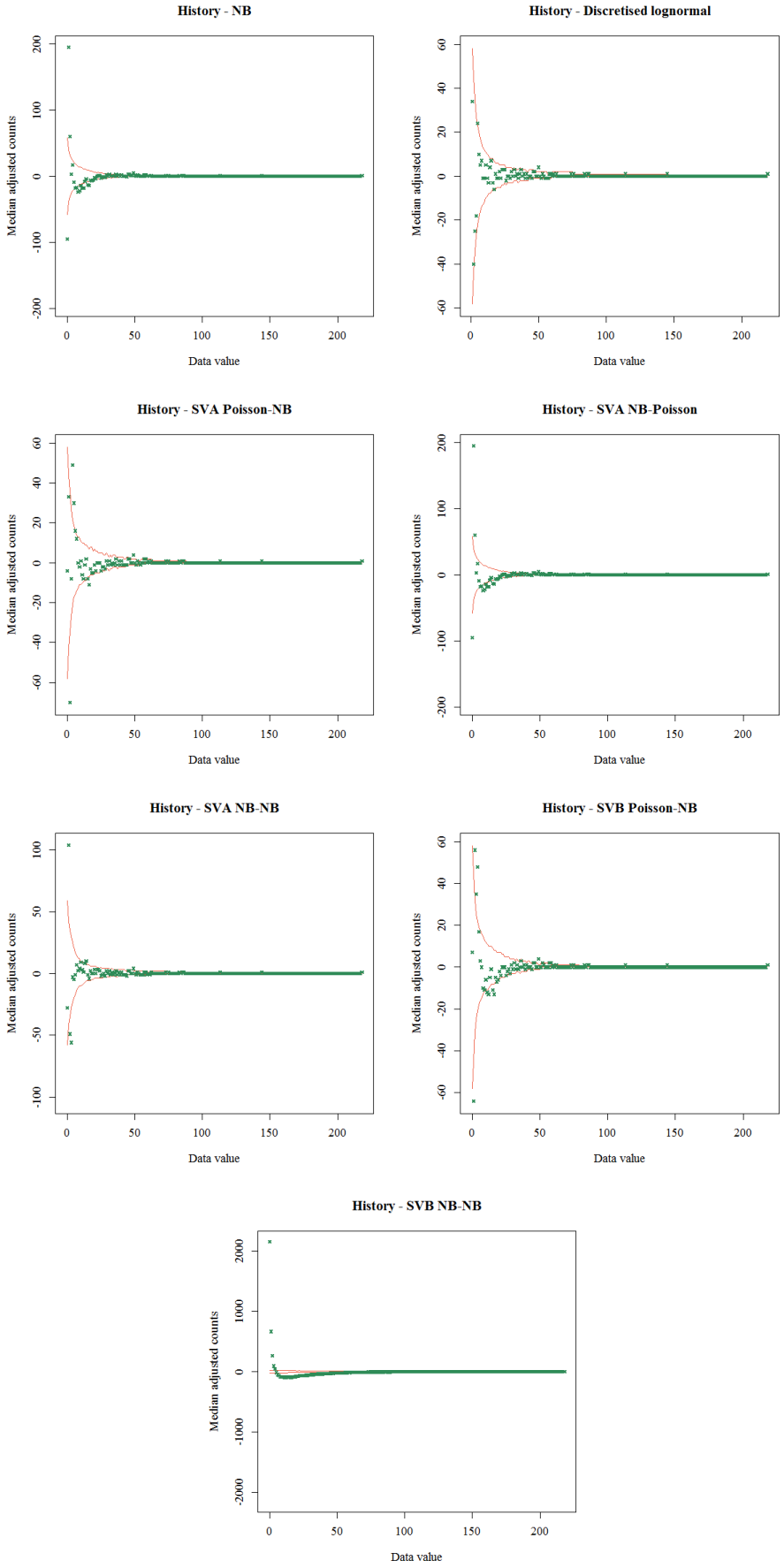

Figure G.6: *Christmas tree plots for History.*

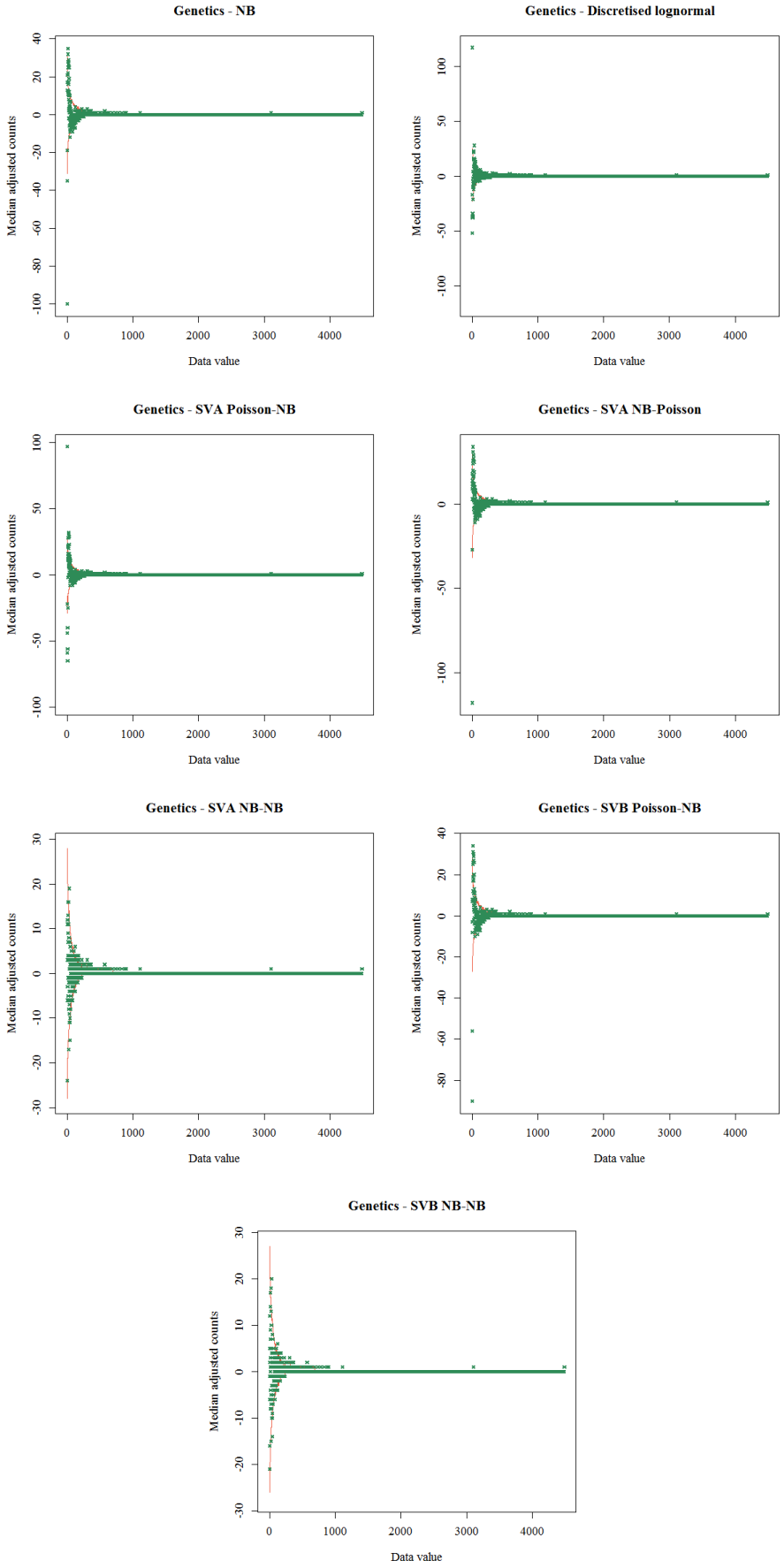

Figure G.7: *Christmas tree plots for Genetics.*

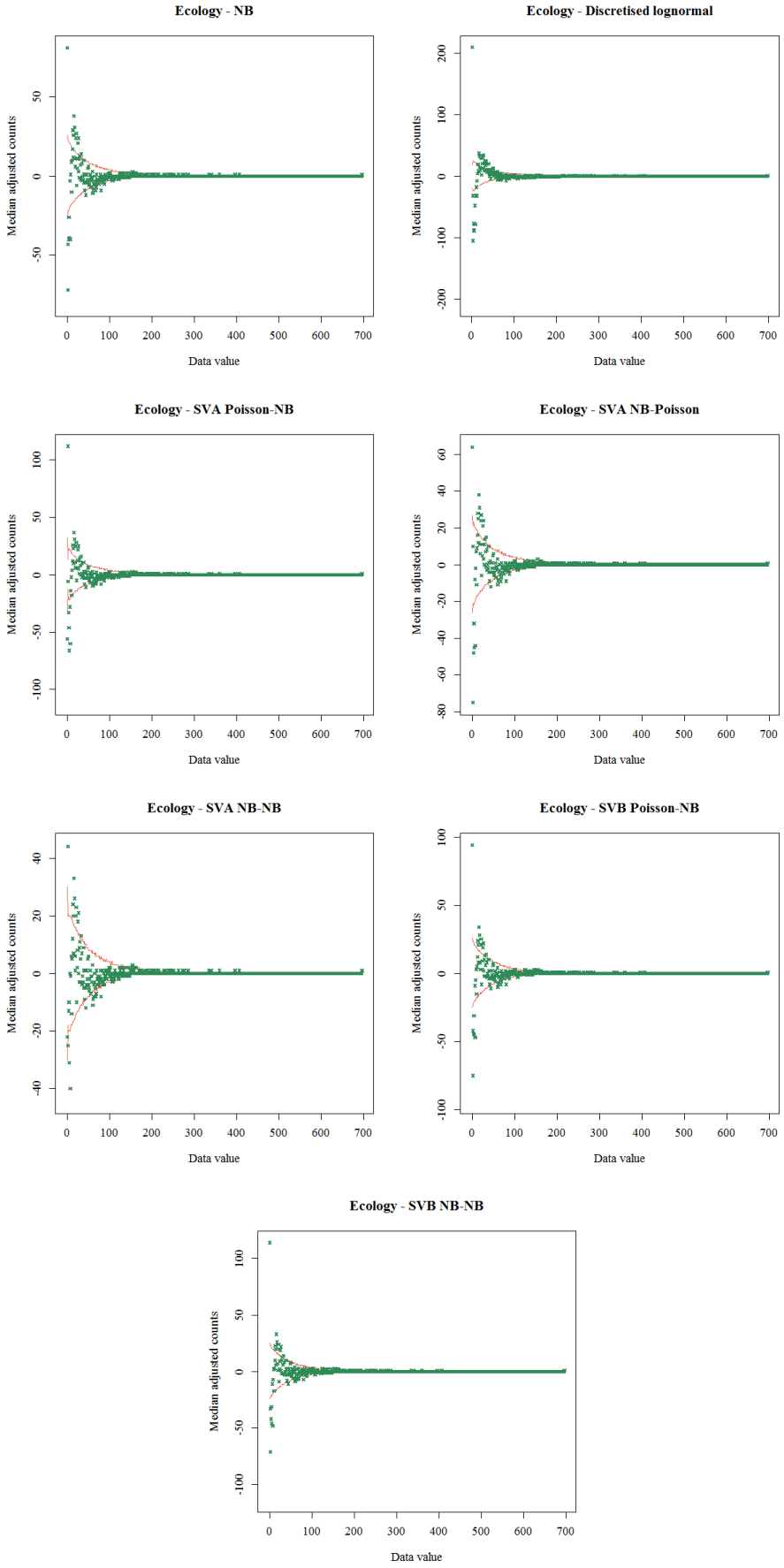

Figure G.8: *Christmas tree plots for Ecology.*

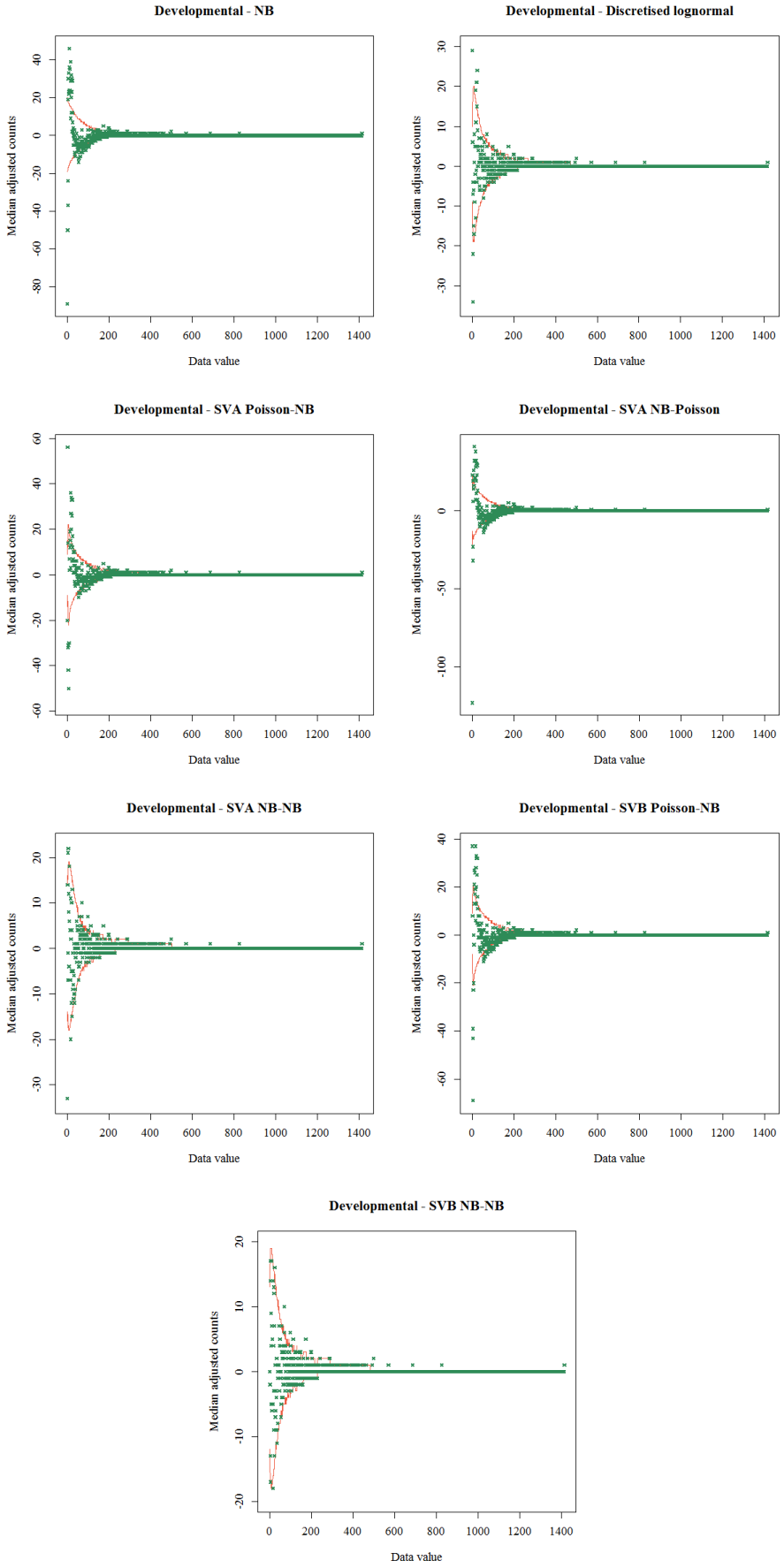

Figure G.9: *Christmas tree plots for Developmental.*

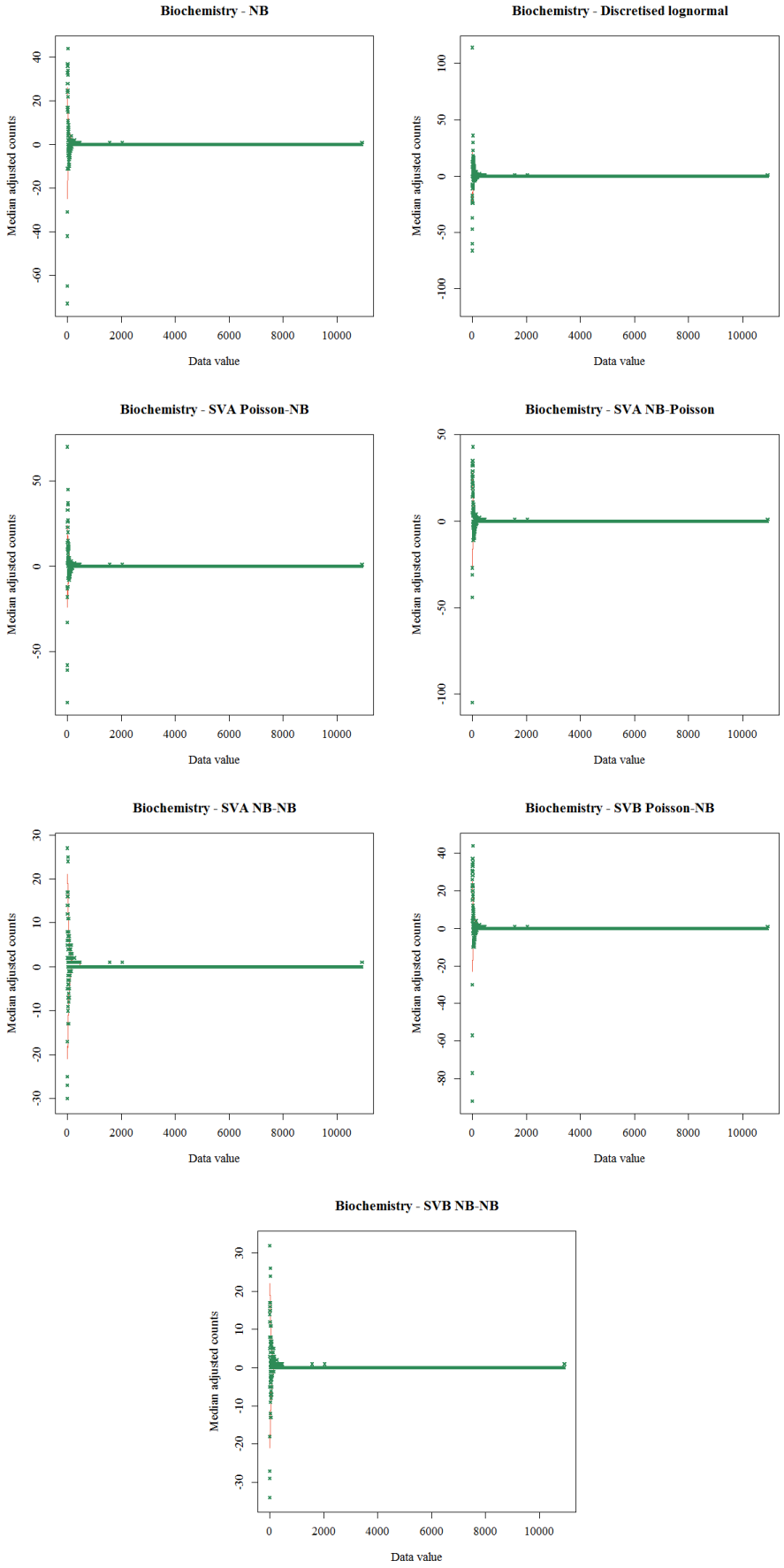

Figure G.10: *Christmas tree plots for Biochemistry.*

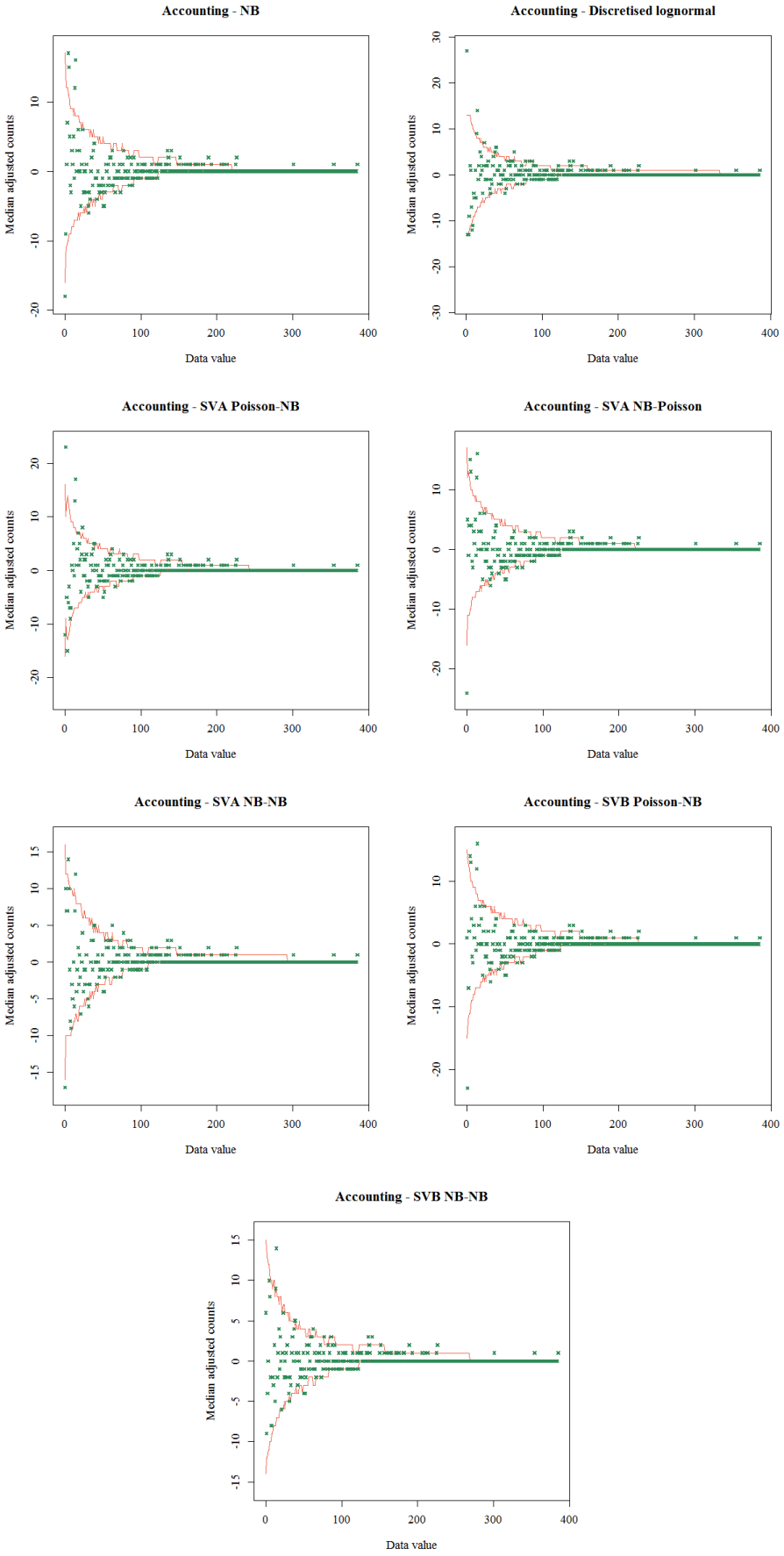

Figure G.11: *Christmas tree plots for Accounting.*

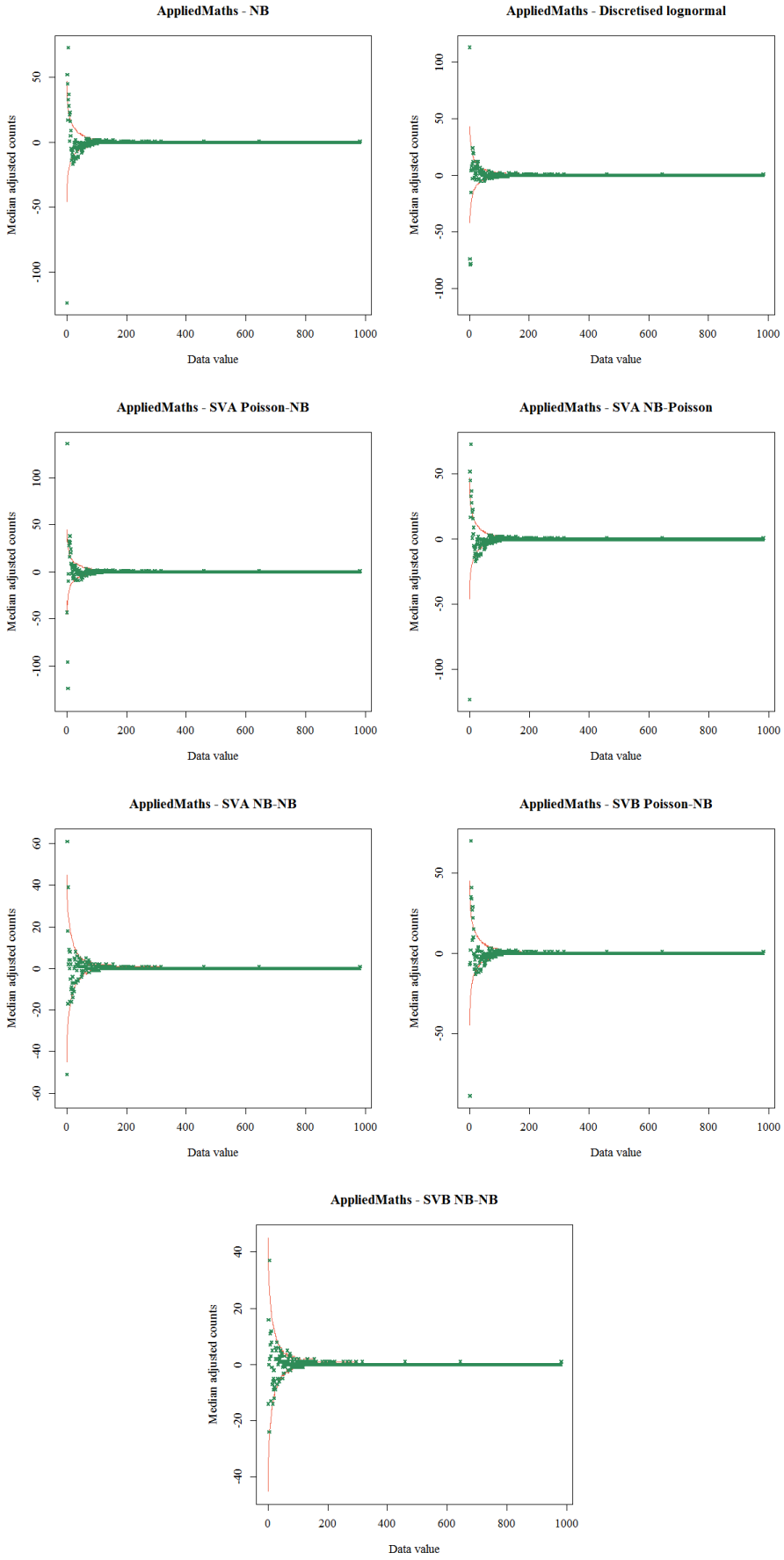

Figure G.12: *Christmas tree plots for AppliedMaths.*

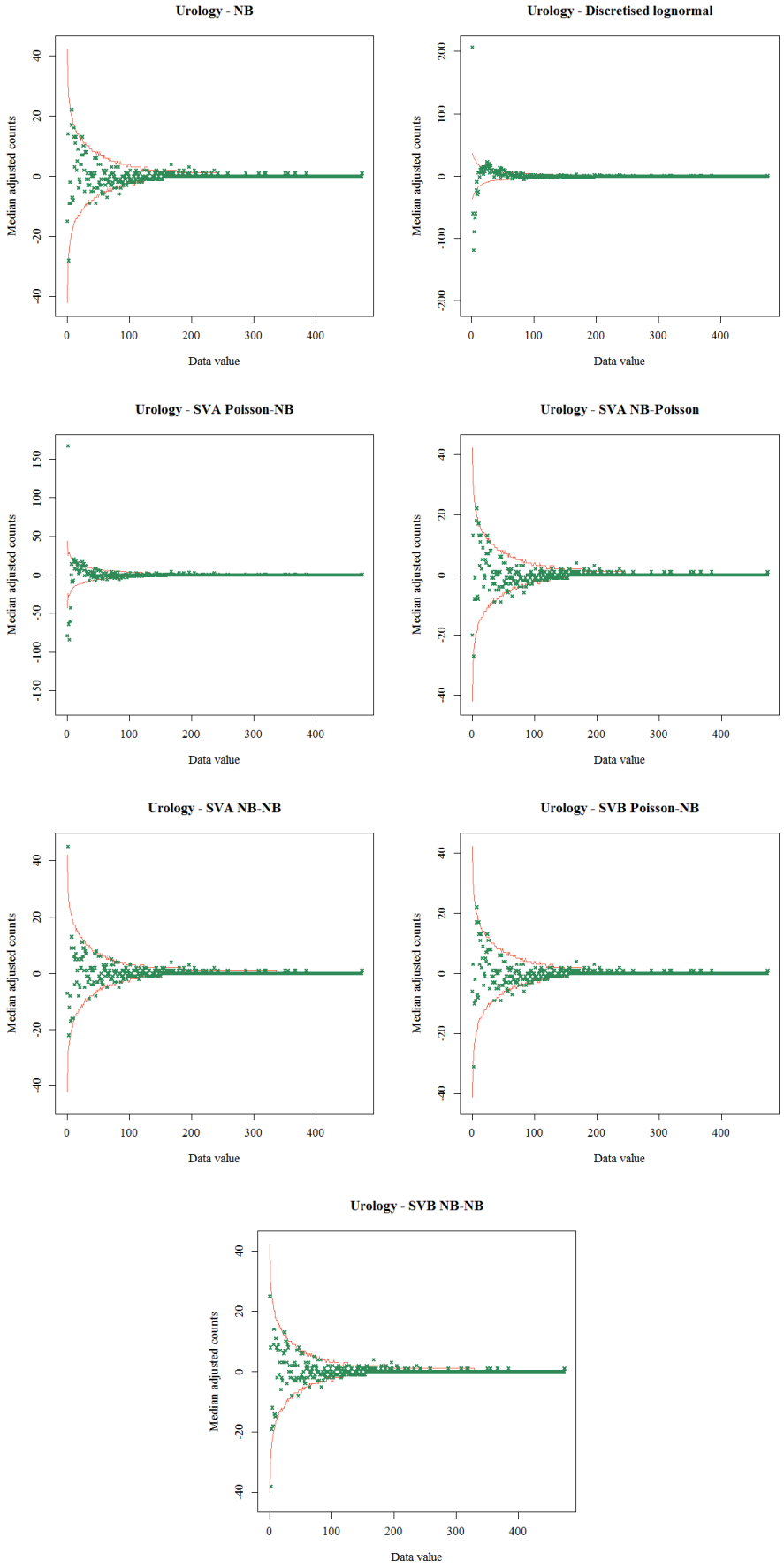

Figure G.13: *Christmas tree plots for Urology.*

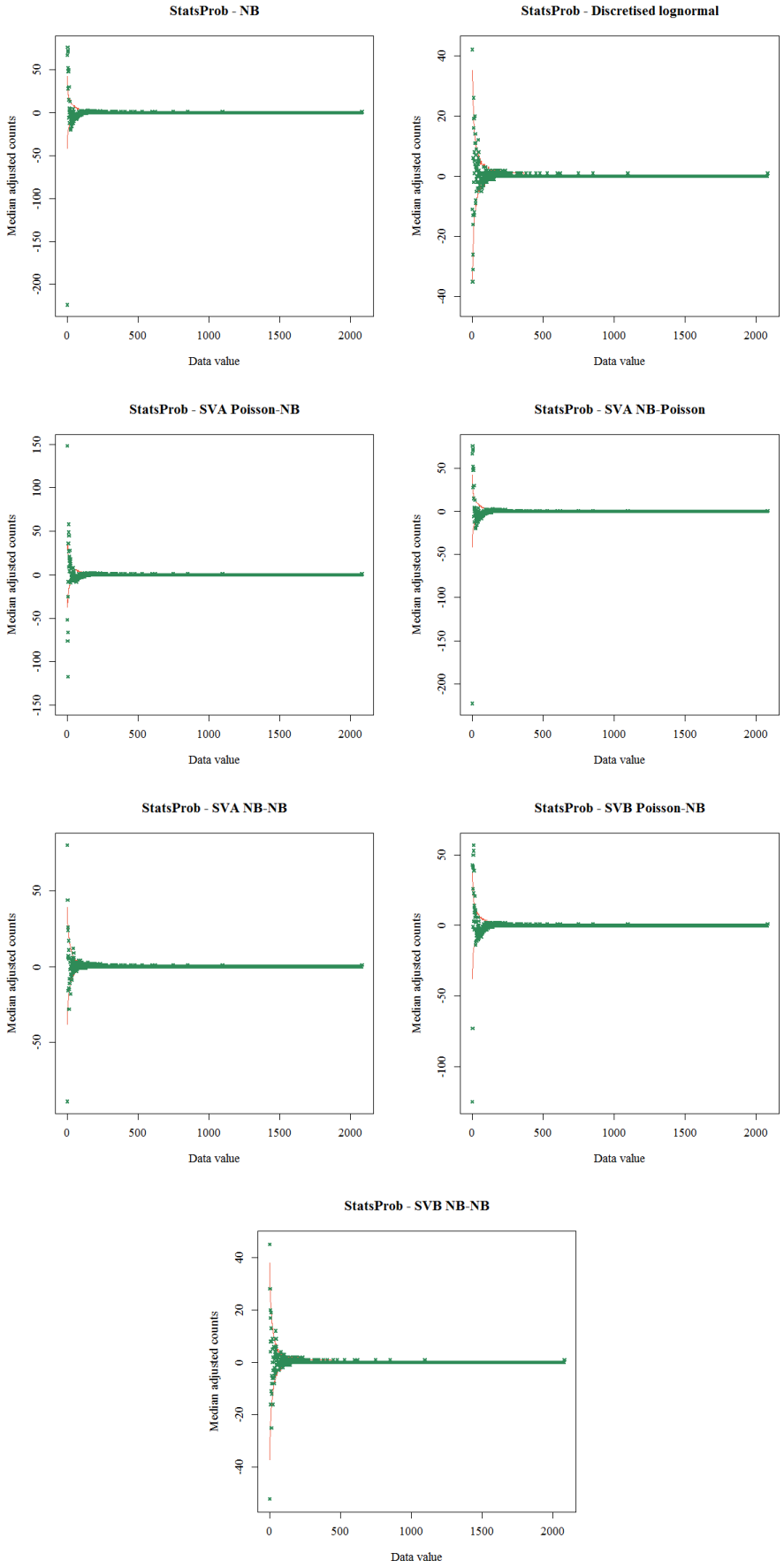

Figure G.14: *Christmas tree plots for StatsProb.*

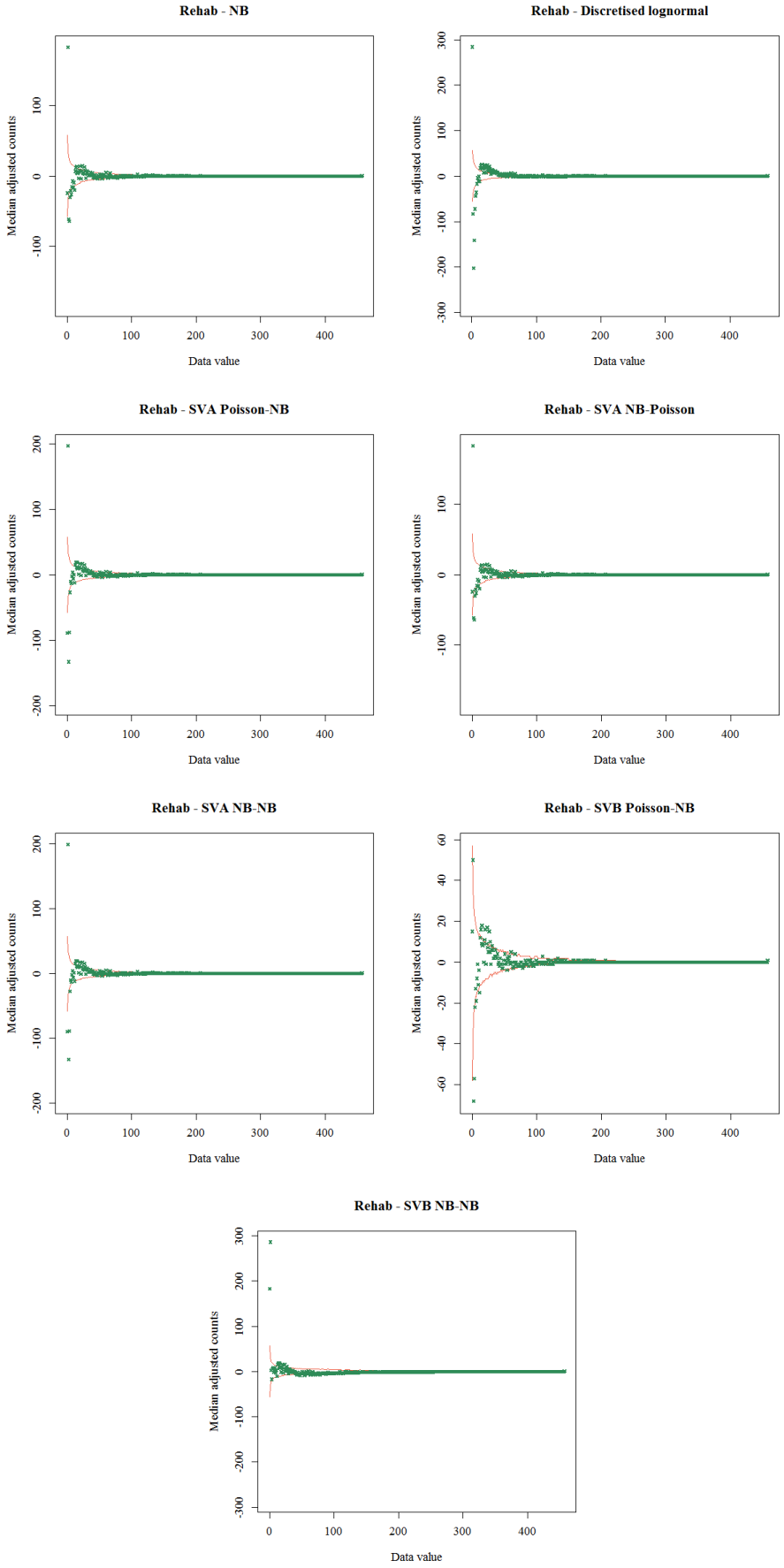

Figure G.15: *Christmas tree plots for Rehab.*

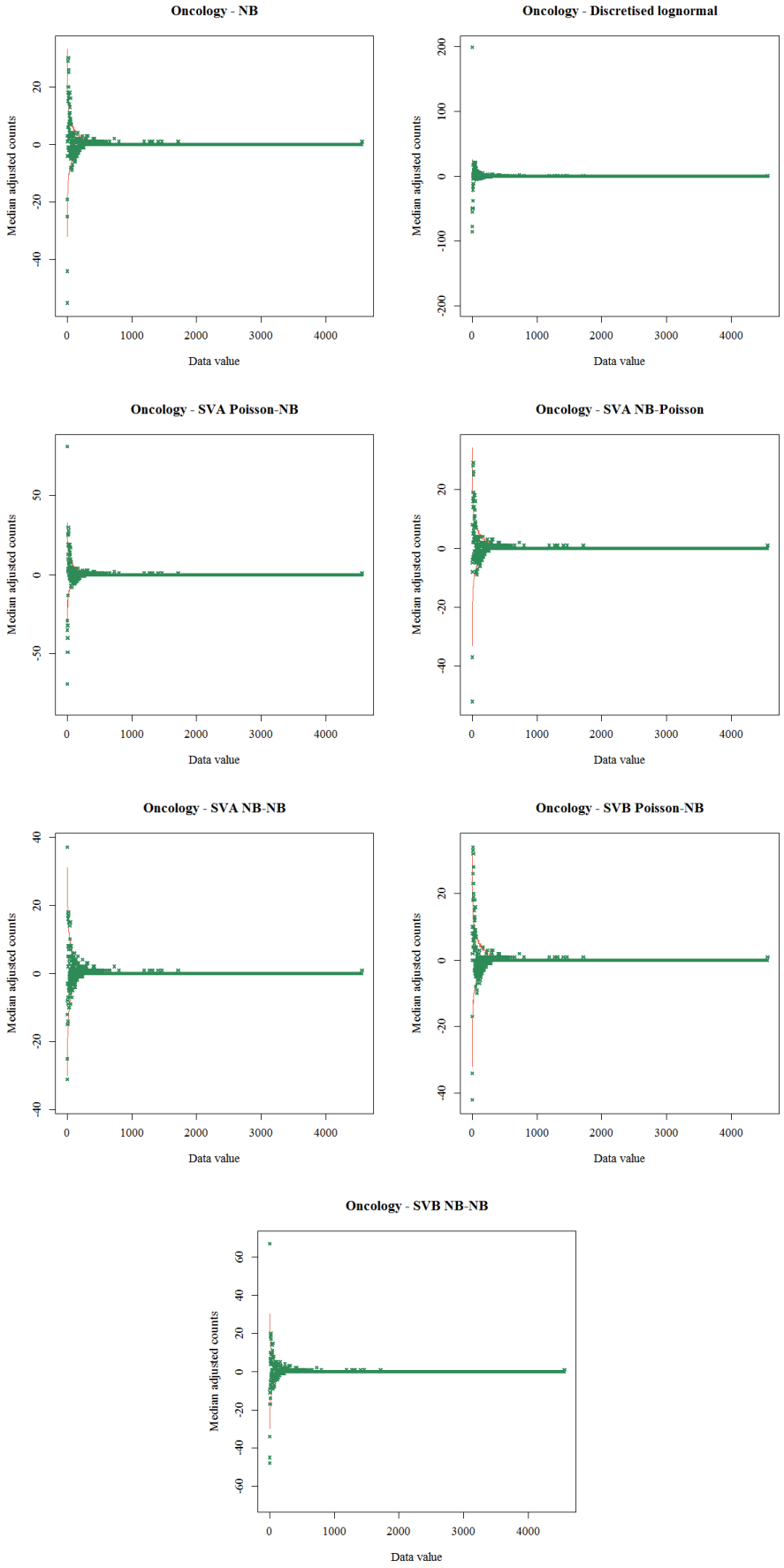

Figure G.16: *Christmas tree plots for Oncology.*

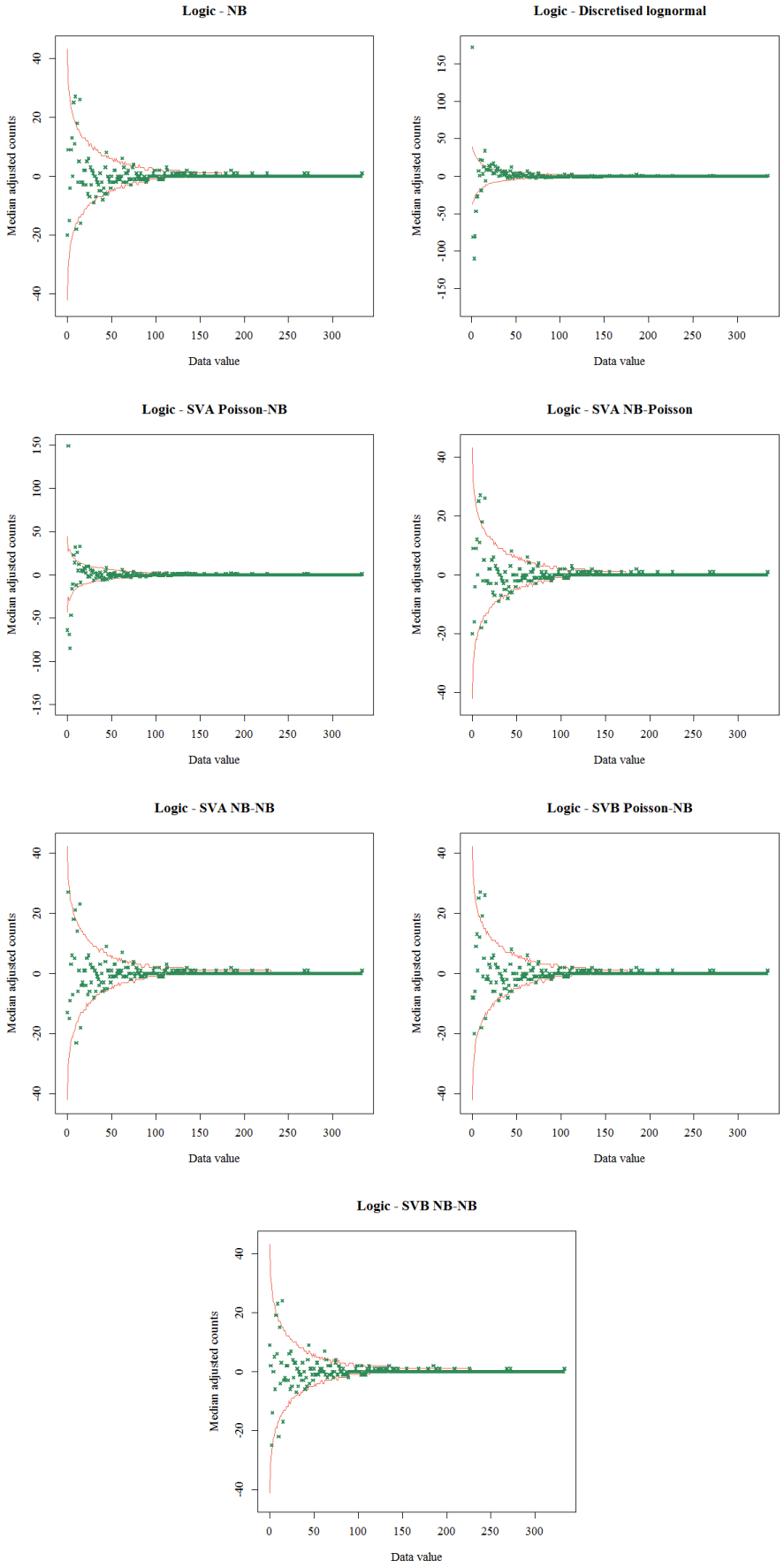

Figure G.17: *Christmas tree plots for Logic.*

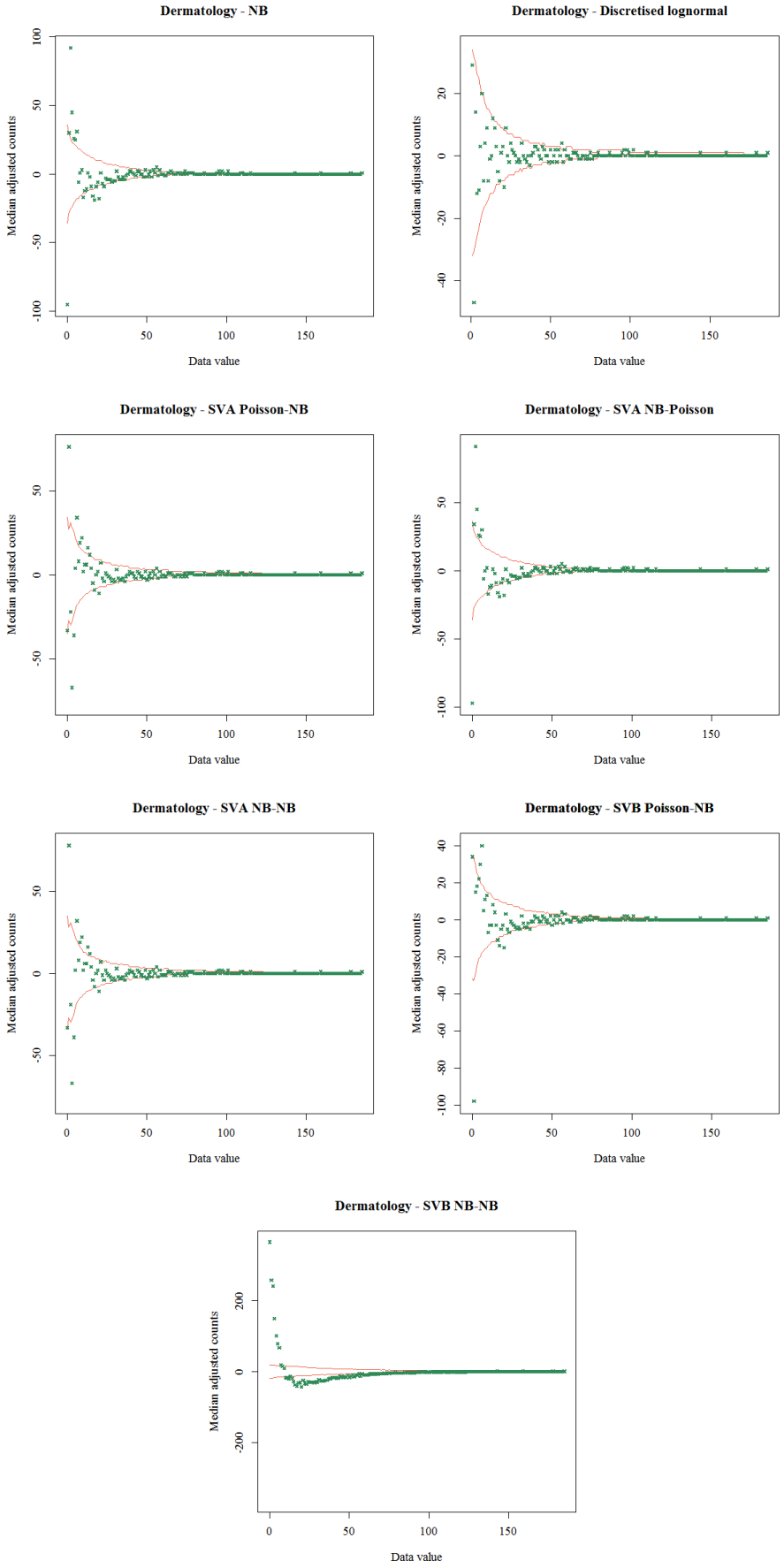

Figure G.18: *Christmas tree plots for Dermatology.*

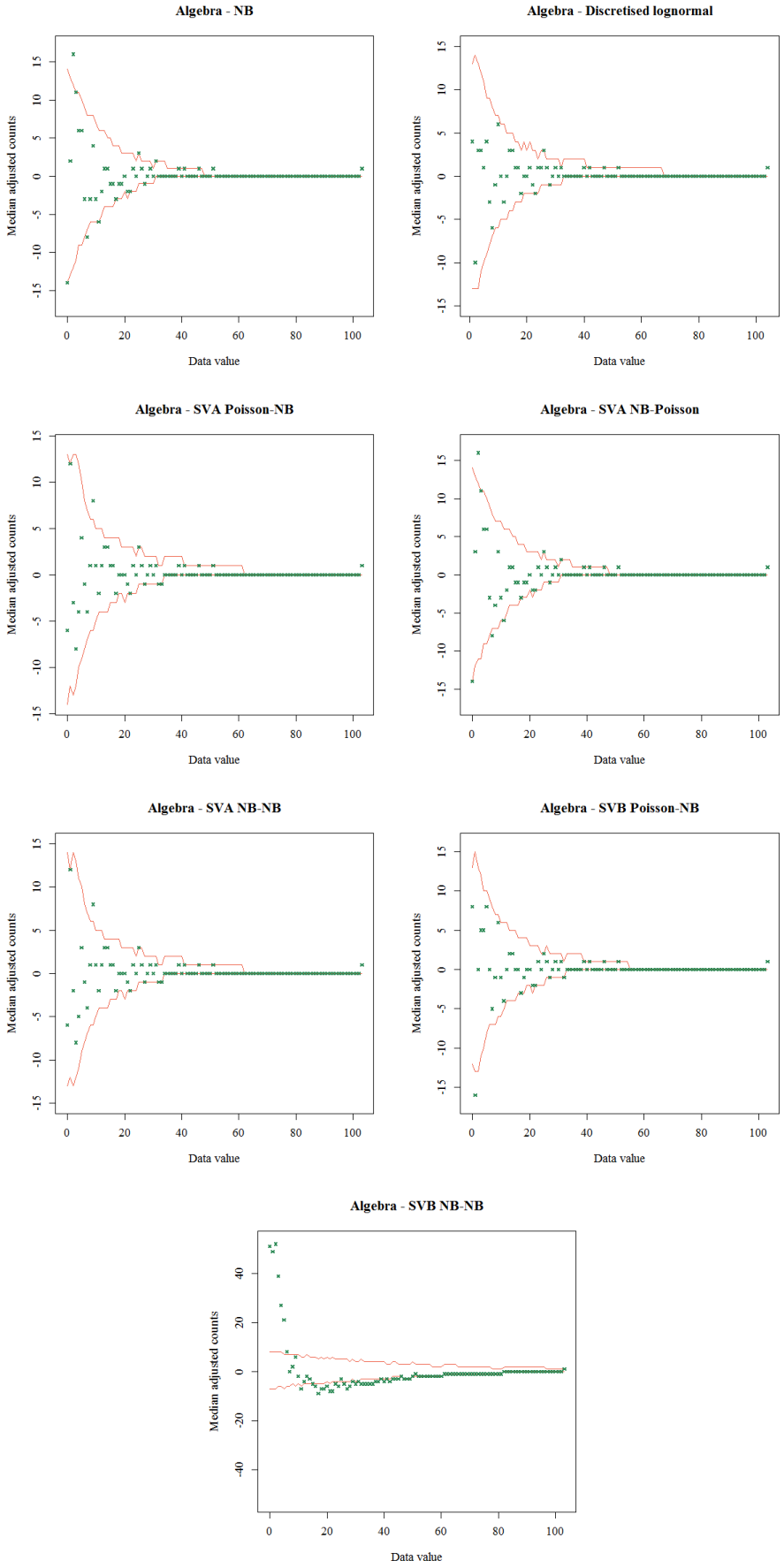

Figure G.19: *Christmas tree plots for Algebra.*

Appendix H

Christmas tree plots for citation analysis with covariates

This section presents the Christmas tree plots for some models fitted to citation data sets in Section 5.4. In all cases, the orange lines denote the 90% fluctuation intervals while the green crosses are the observed median adjusted counts.

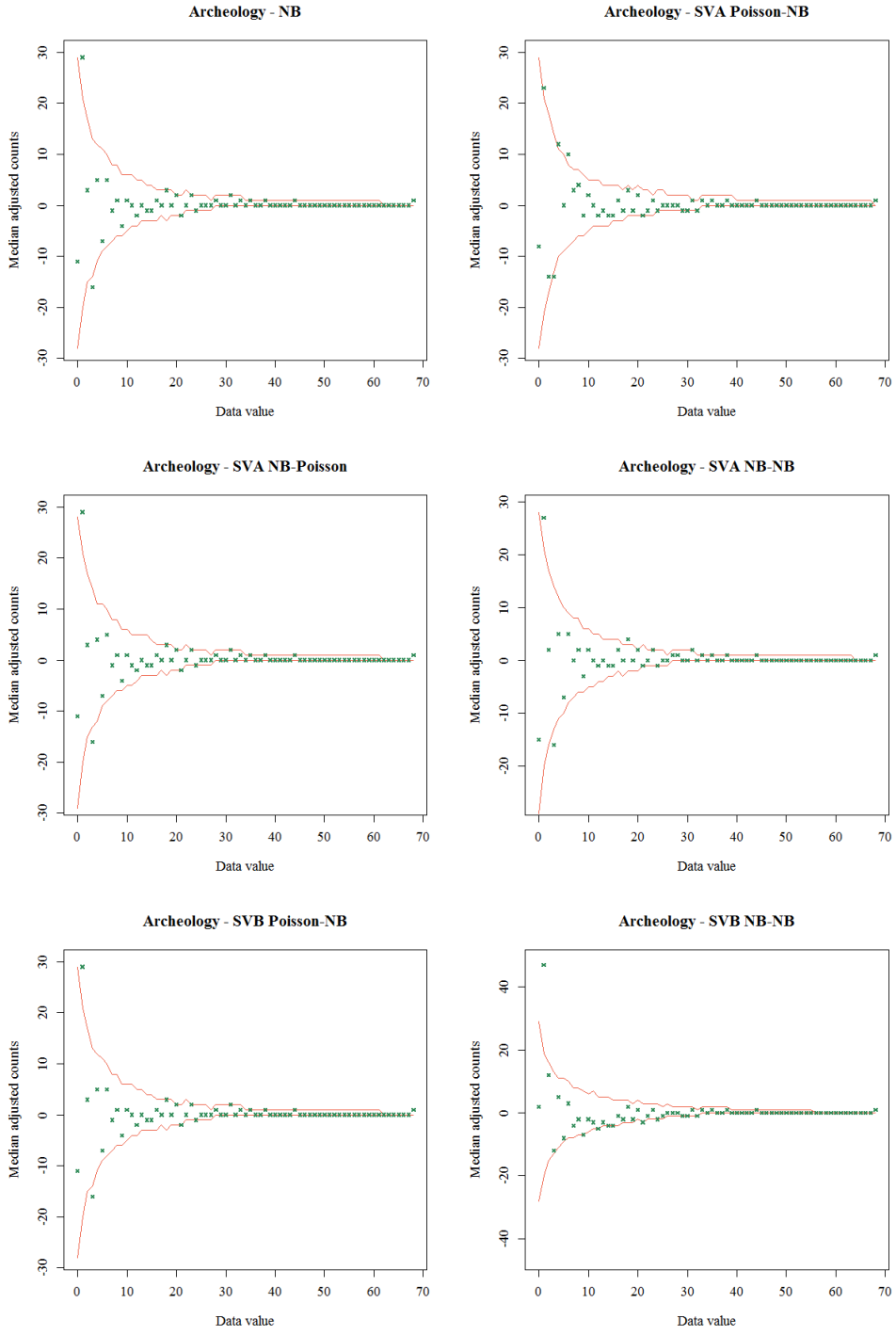

Figure H.1: *Christmas tree plots for Archeology.*

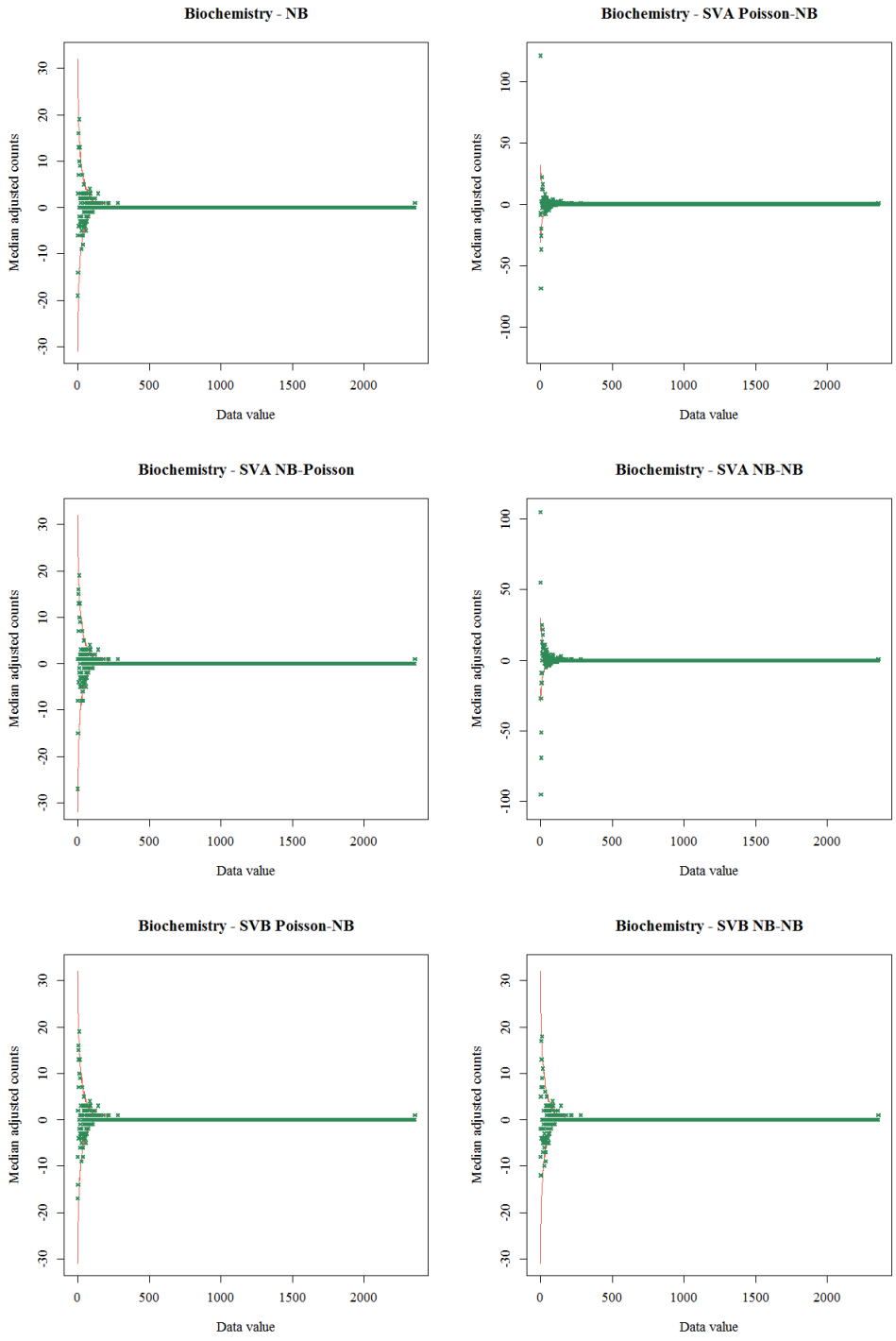

Figure H.2: *Christmas tree plots for Biochemistry.*

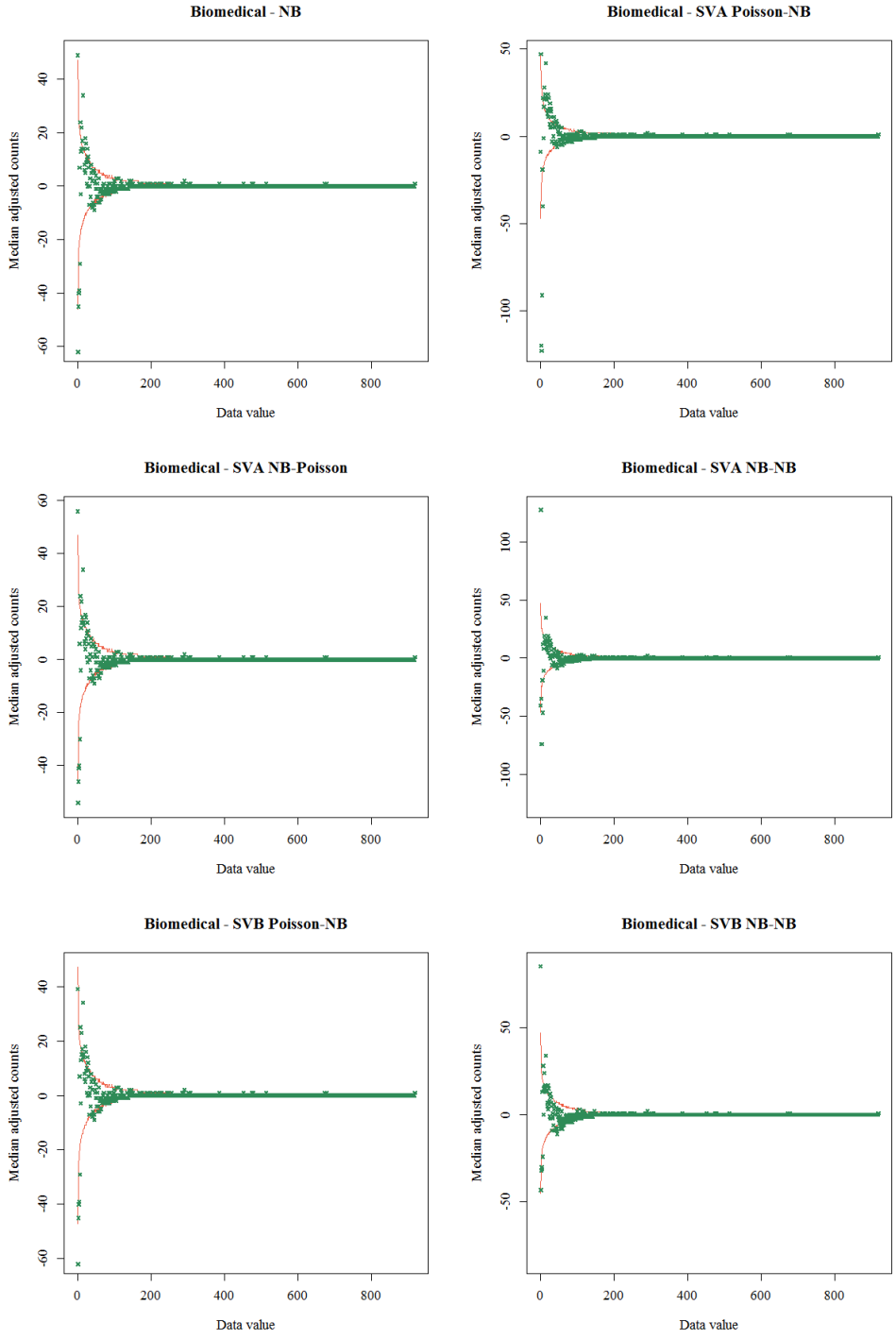

Figure H.3: *Christmas tree plots for Biomedical Engineering.*

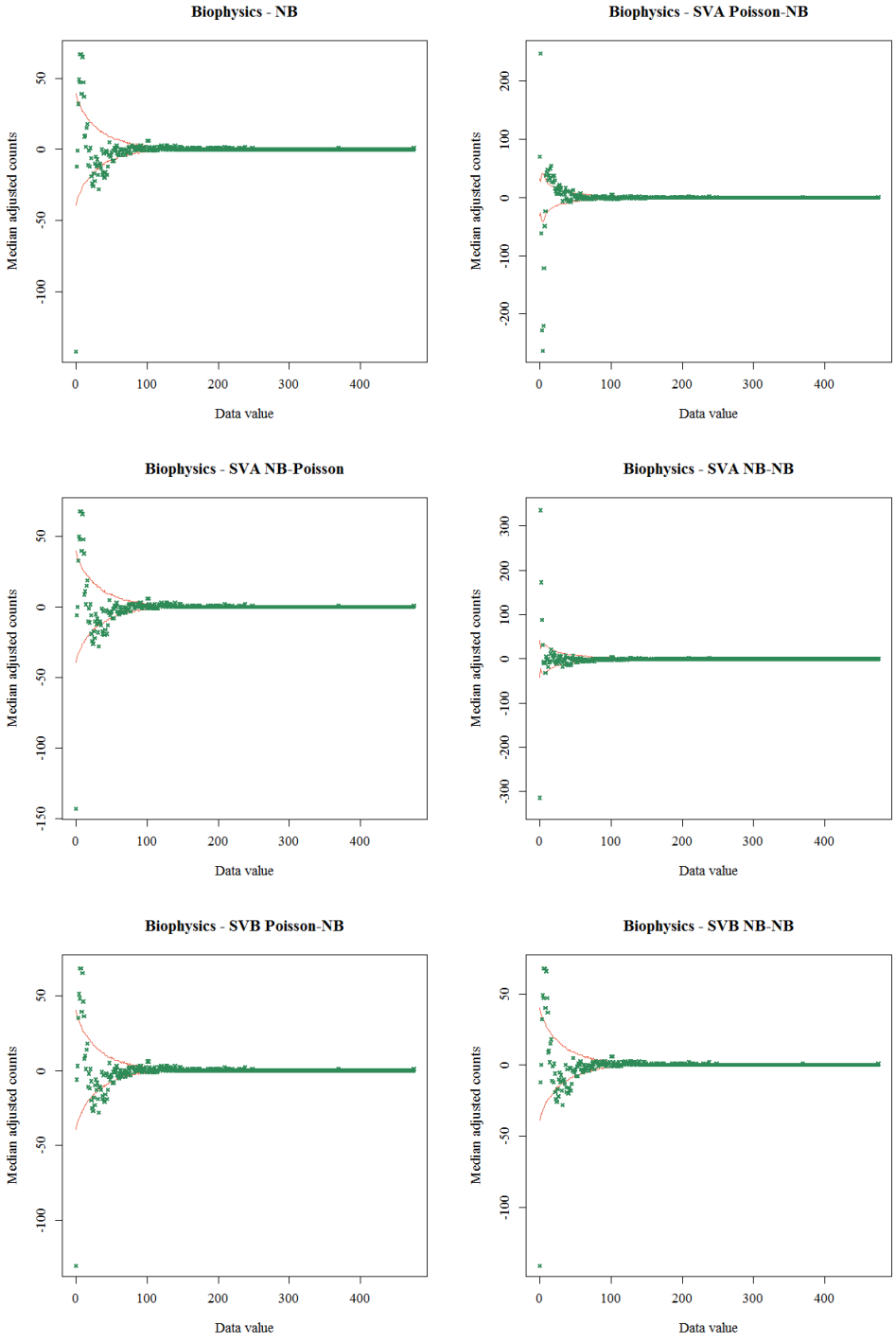

Figure H.4: *Christmas tree plots for Biophysics.*

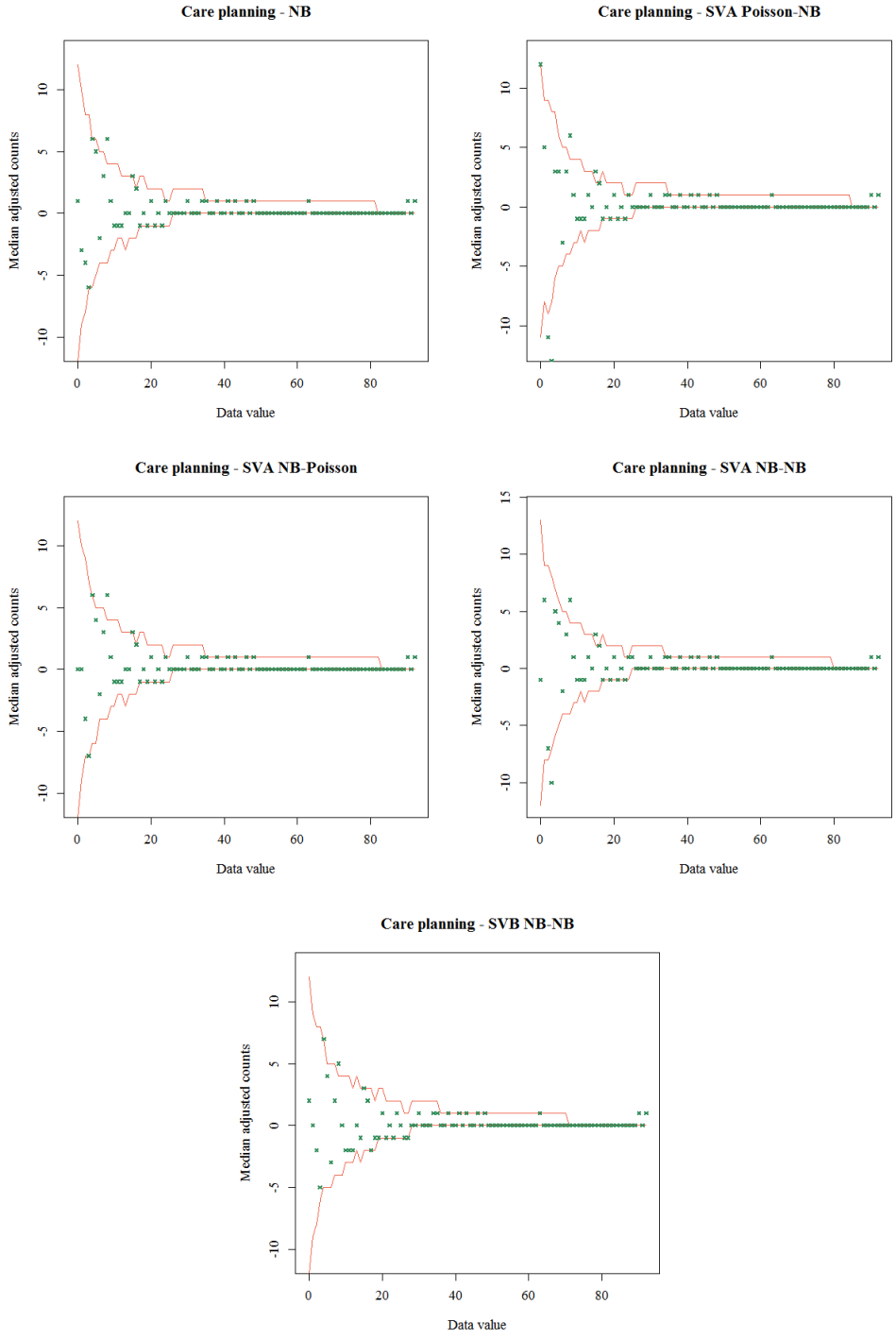

Figure H.5: *Christmas tree plots for Care Planning.*

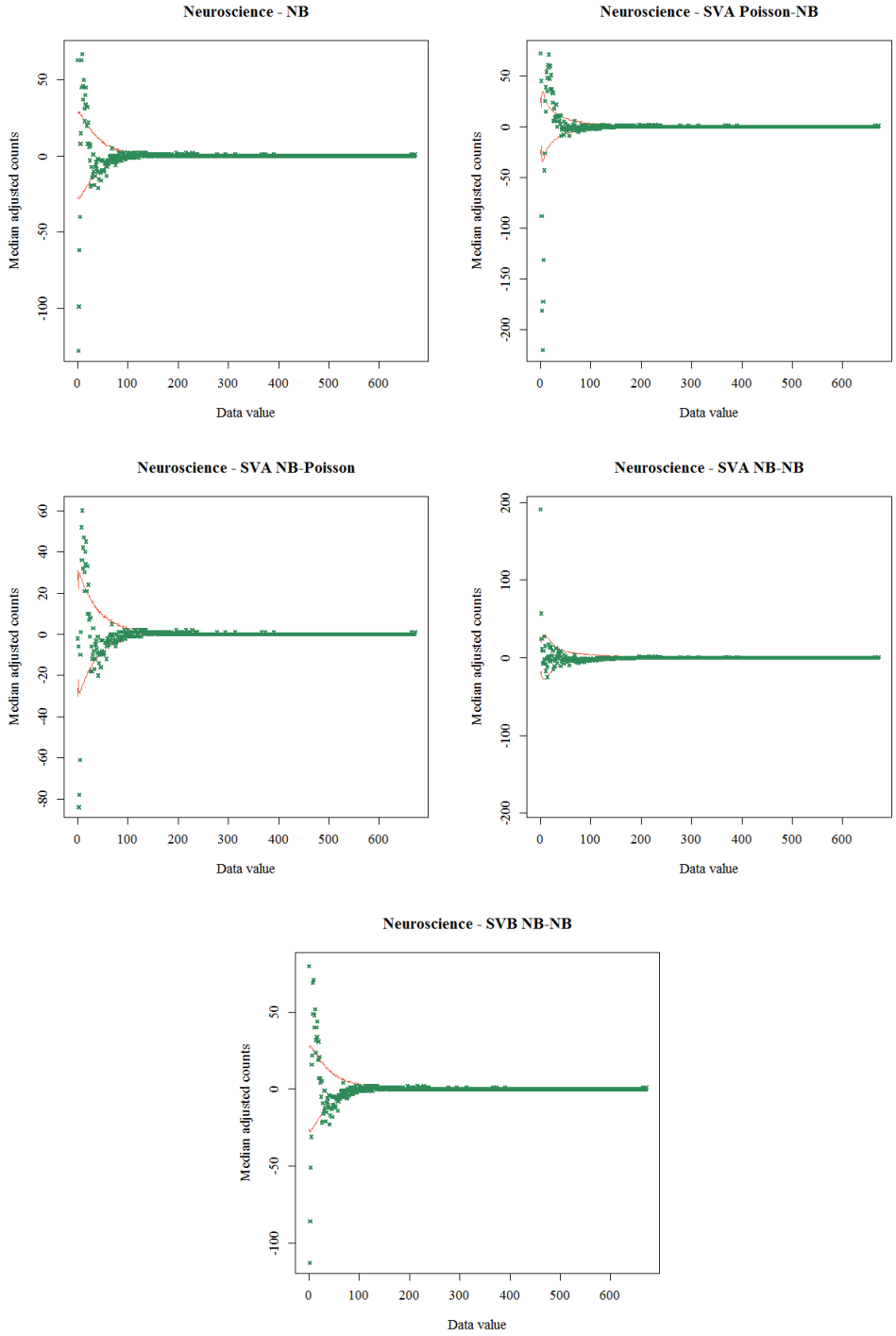

Figure H.6: *Christmas tree plots for Cellular and Molecular Neuroscience.*

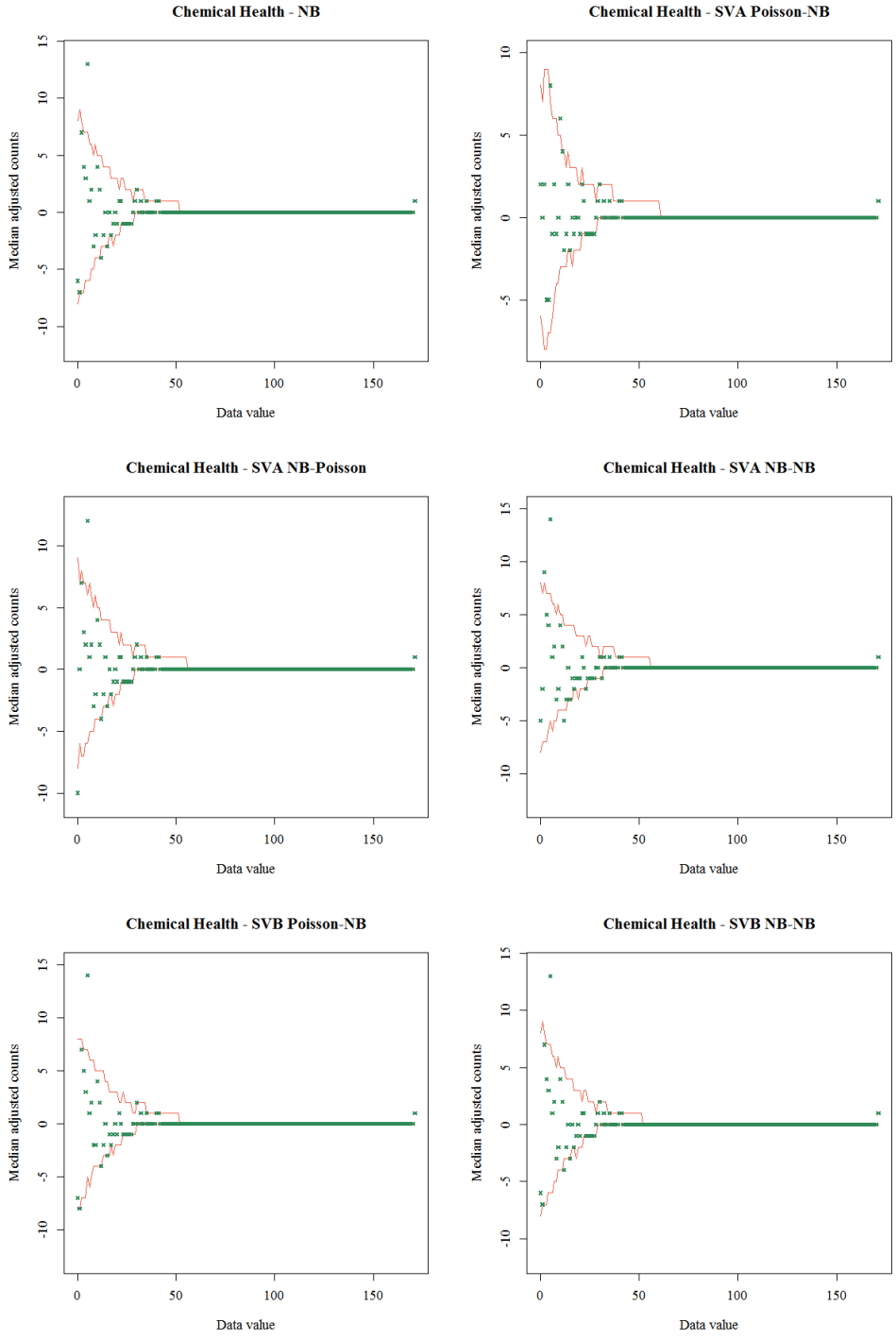

Figure H.7: *Christmas tree plots for Chemical Health and Safety.*

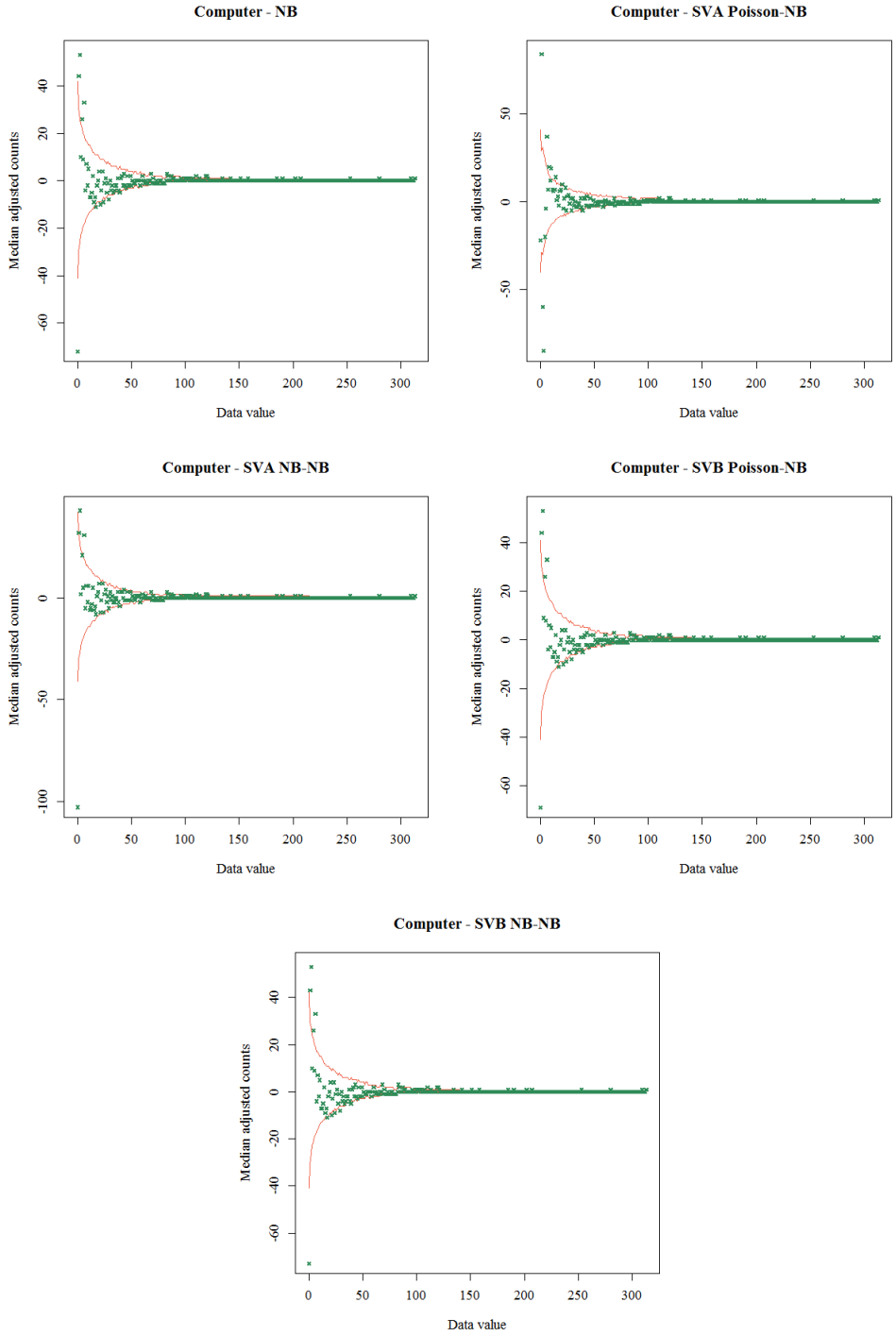

Figure H.8: *Christmas tree plots for Computer Graphics and Computer Aided Design.*

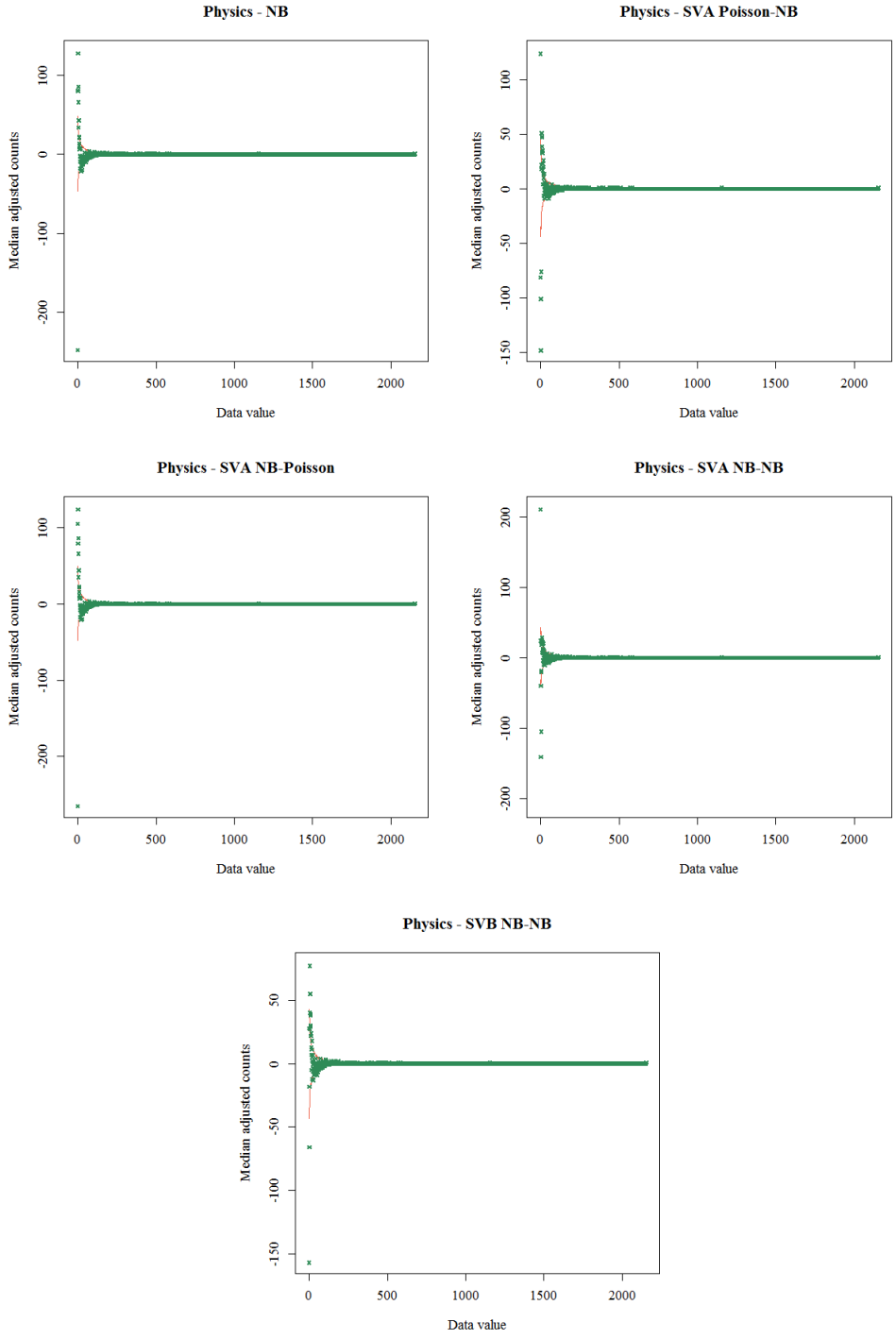

Figure H.9: *Christmas tree plots for Condensed Matter Physics.*

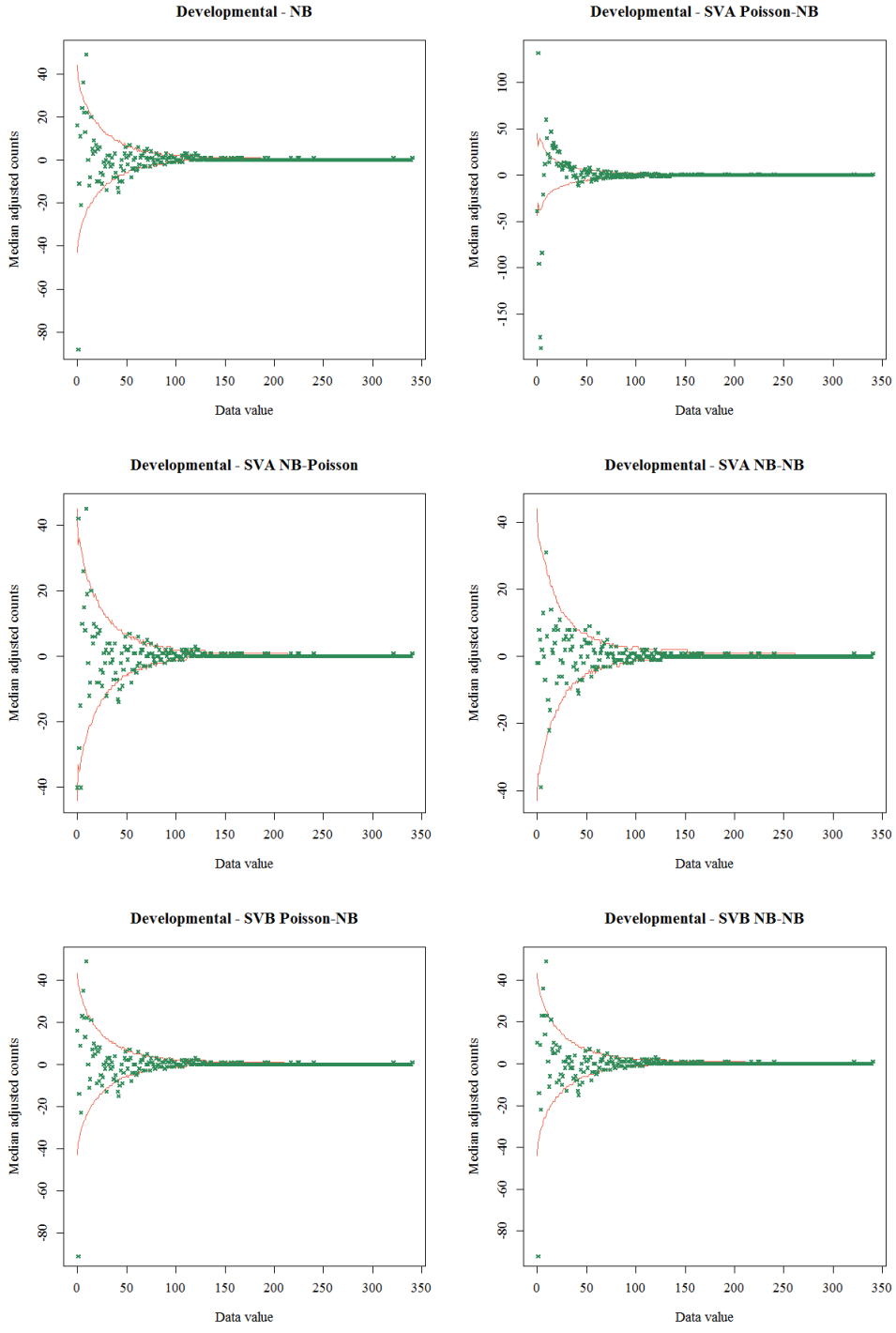

Figure H.10: *Christmas tree plots for Developmental and Educational Psychology.*

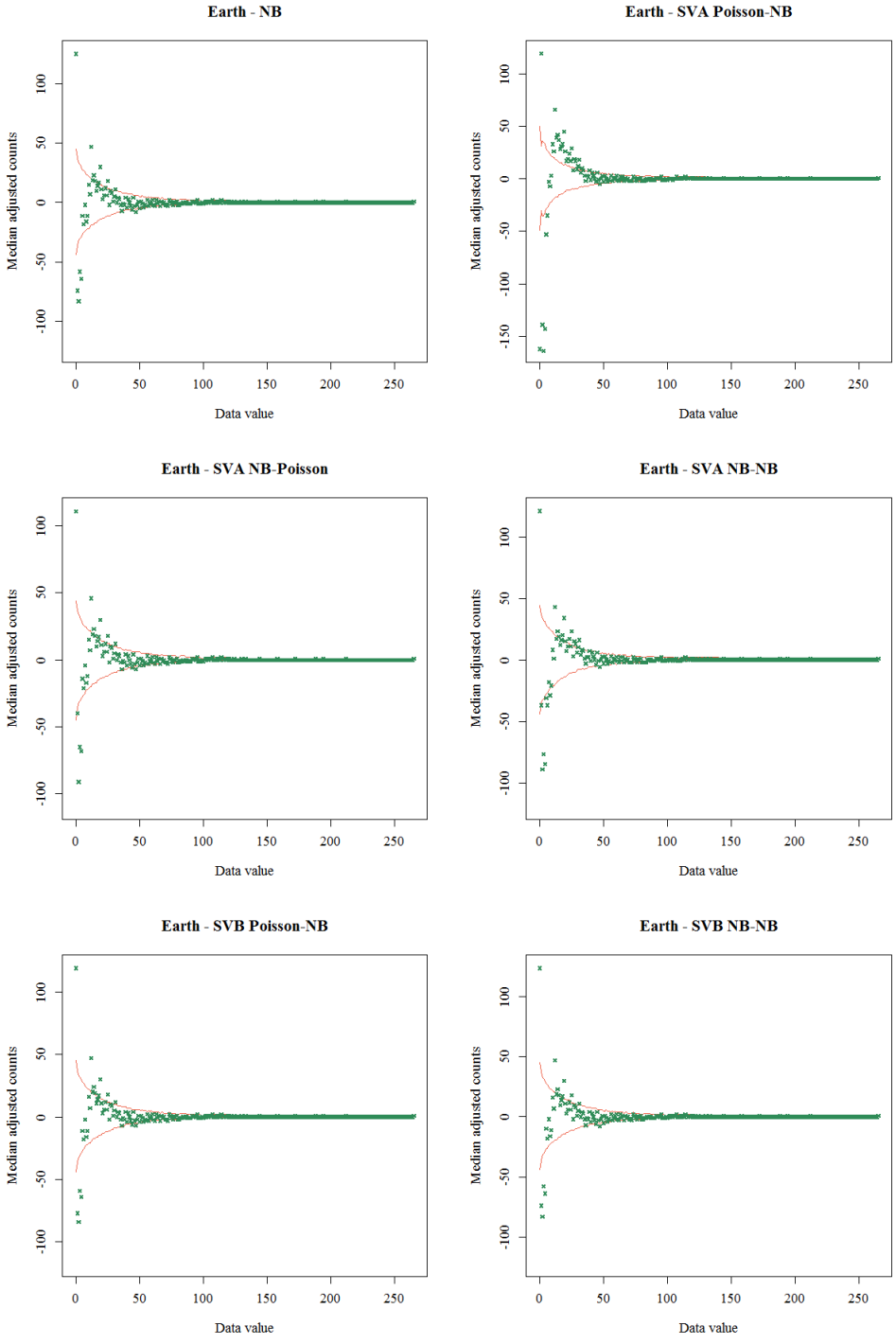

Figure H.11: *Christmas tree plots for Earth Surface Processes.*

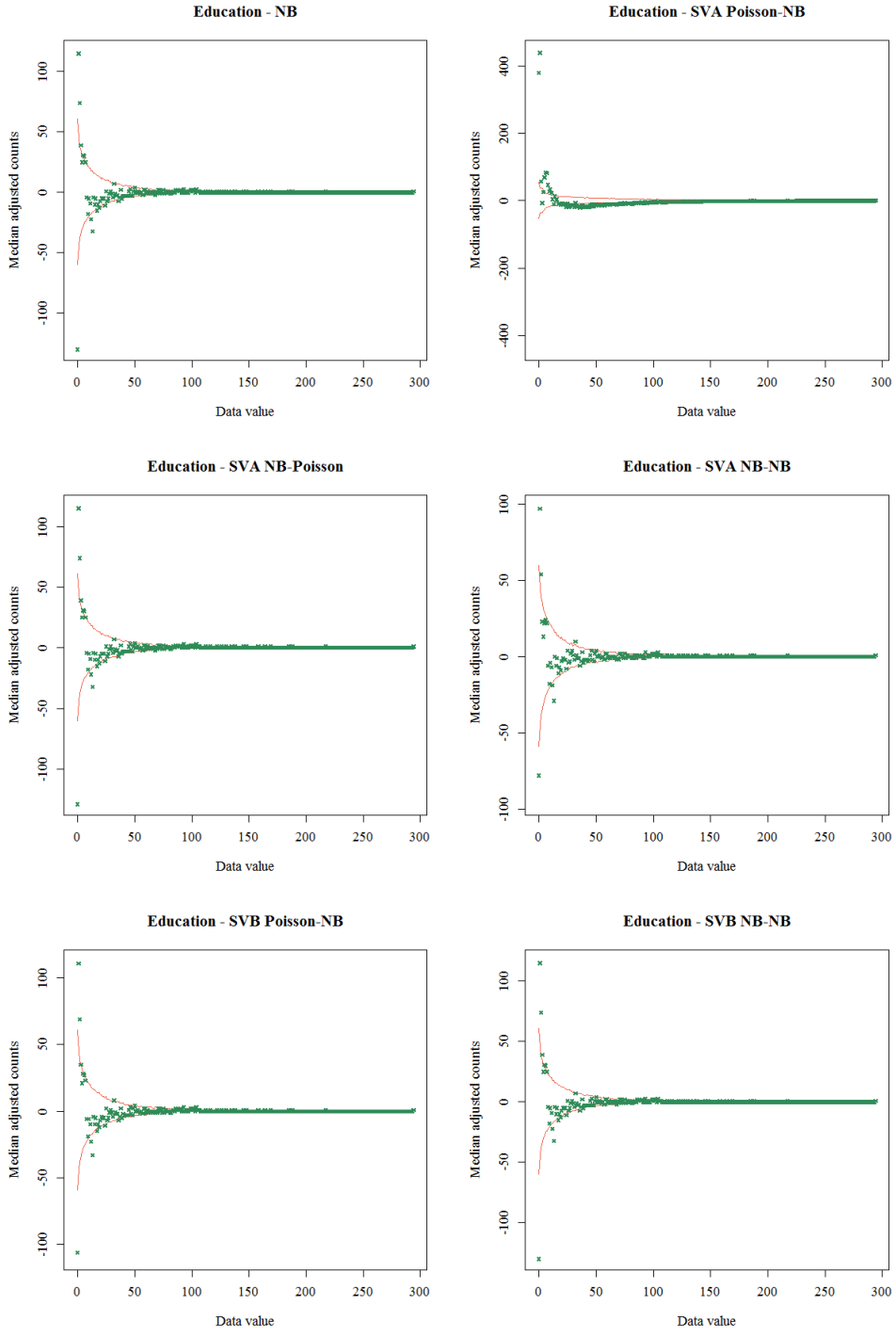

Figure H.12: *Christmas tree plots for Education.*

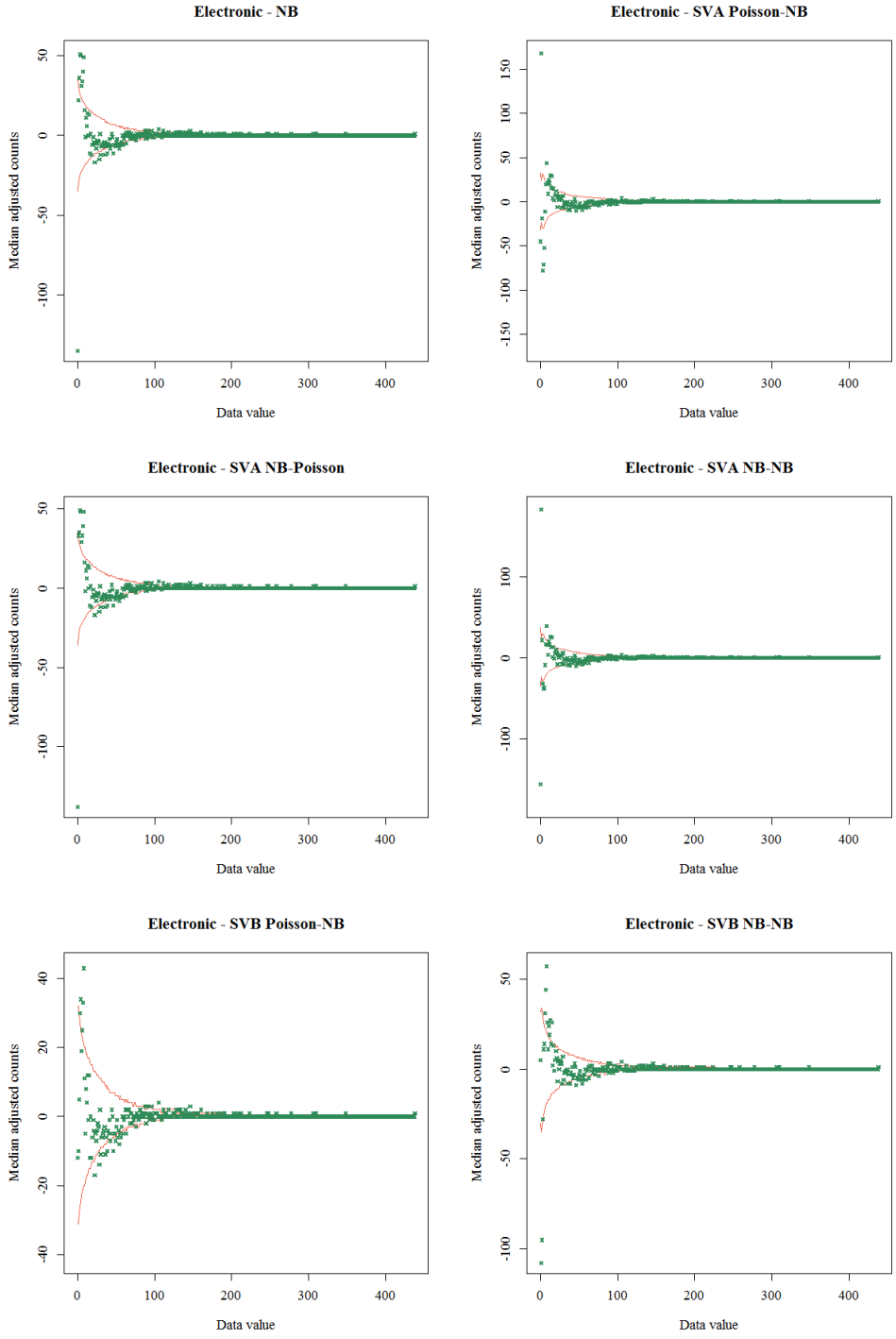

Figure H.13: *Christmas tree plots for Electronic Optical and Magnetic Materials.*

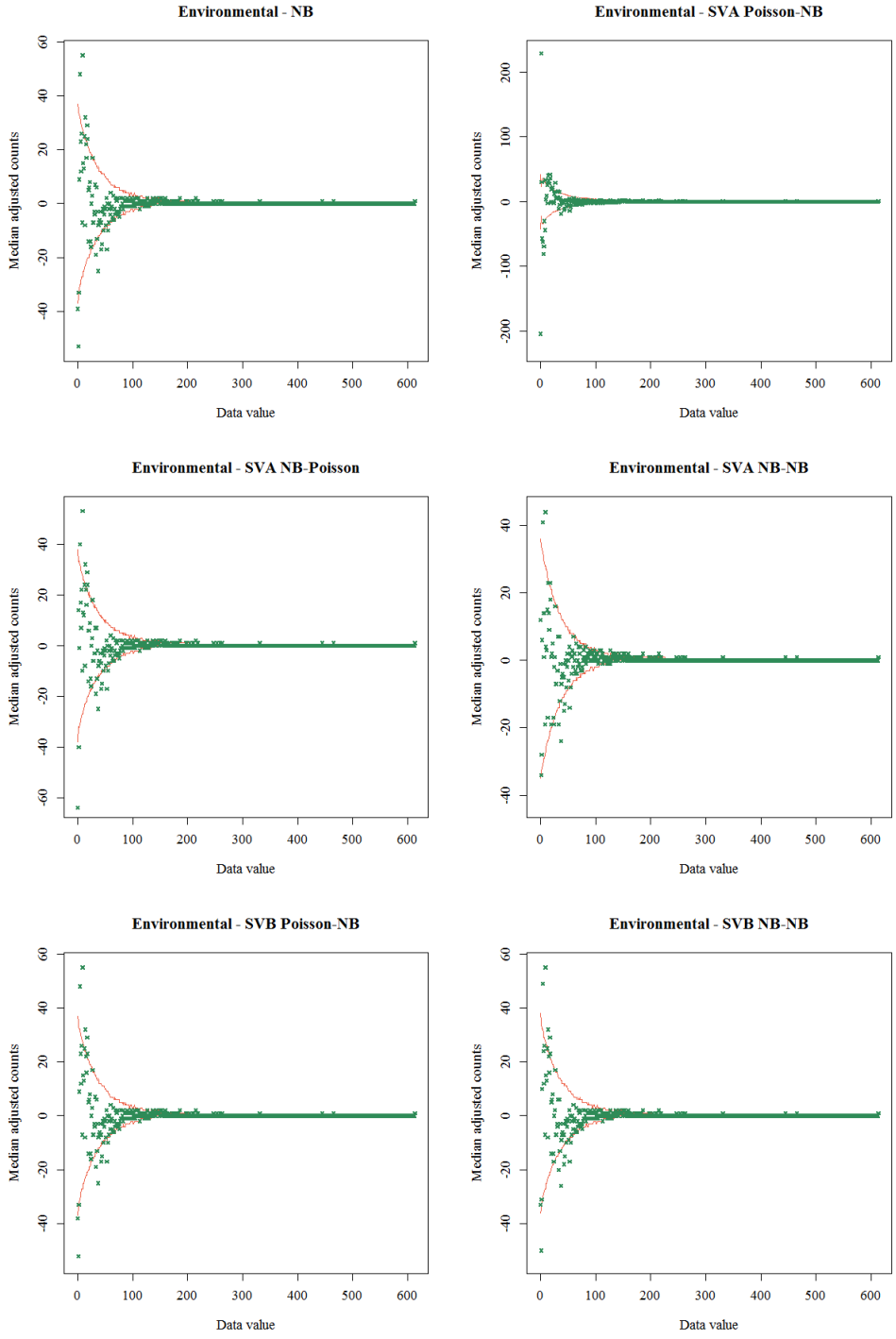

Figure H.14: *Christmas tree plots for Environmental Chemistry.*

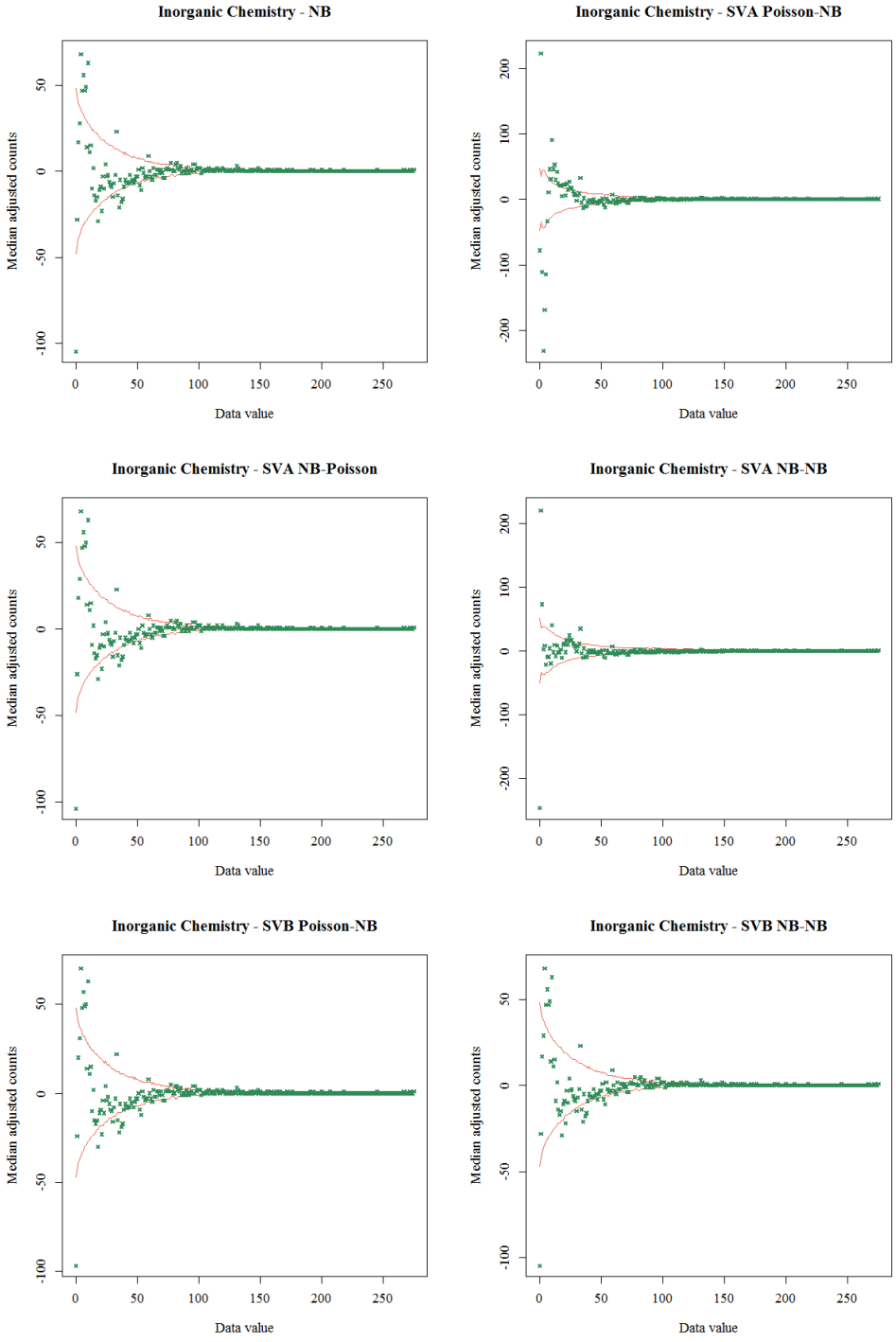

Figure H.15: *Christmas tree plots for Inorganic Chemistry.*

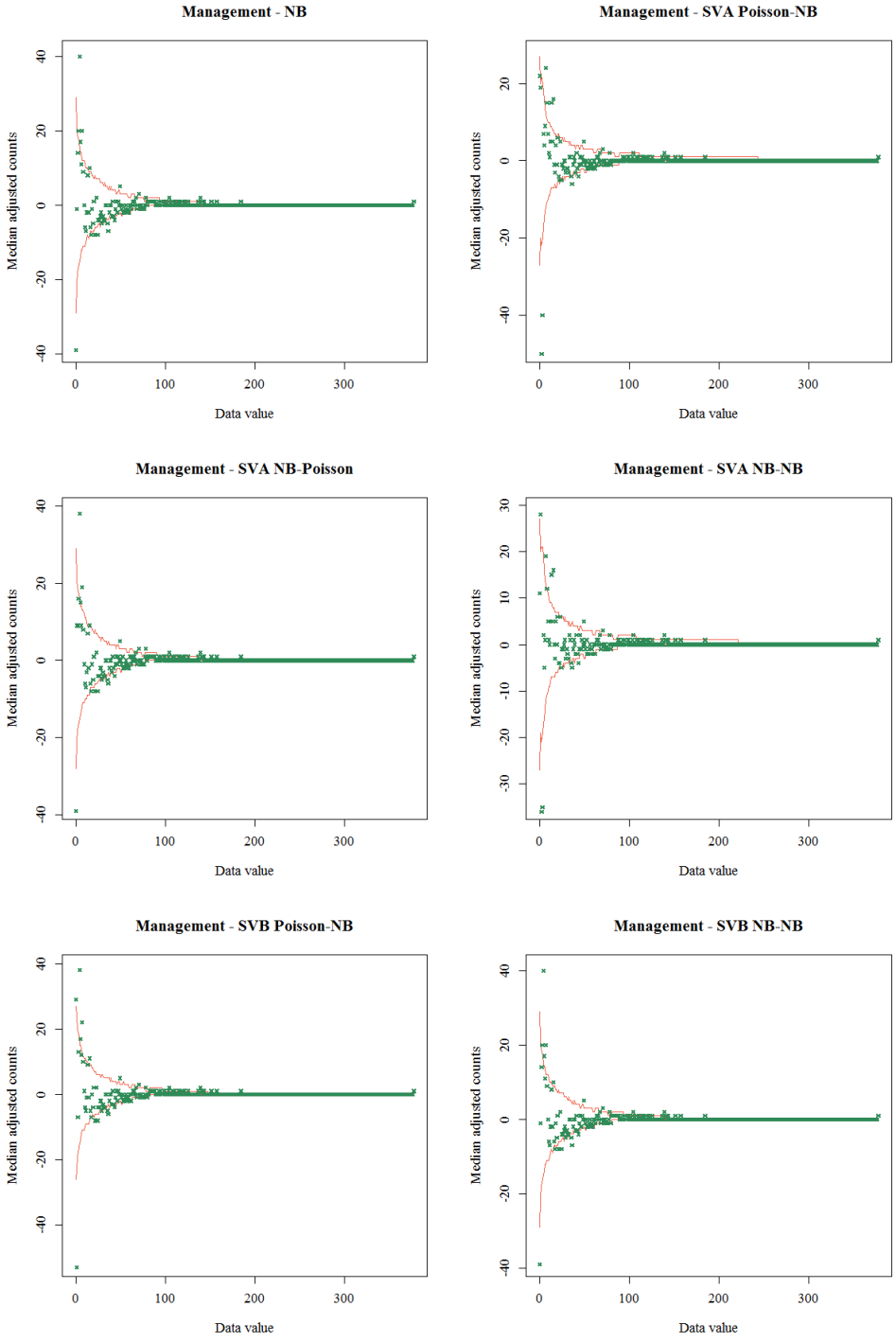

Figure H.16: *Christmas tree plots for Management Information Systems.*

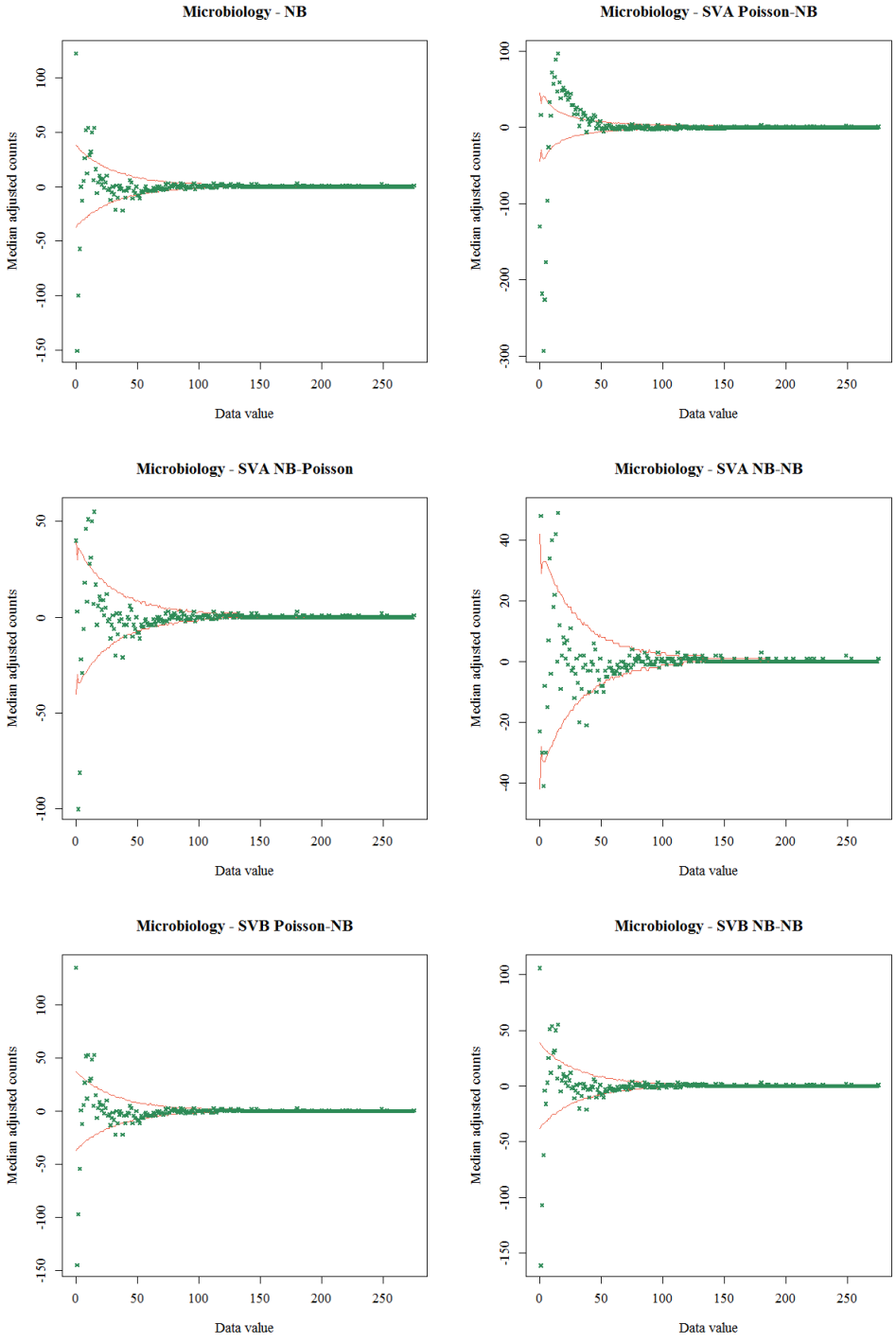

Figure H.17: *Christmas tree plots for Microbiology.*

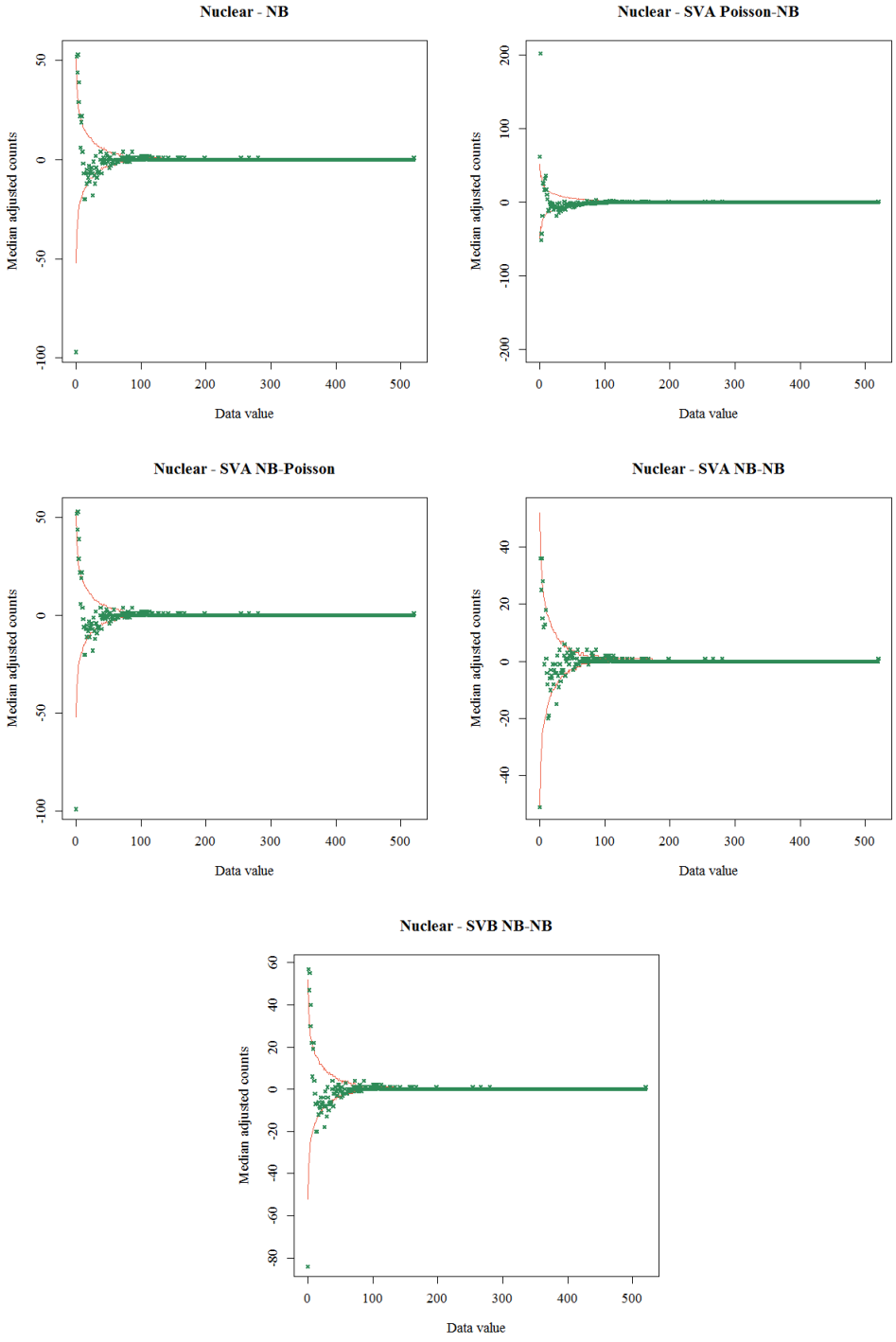

Figure H.18: *Christmas tree plots for Nuclear Energy and Engineering.*

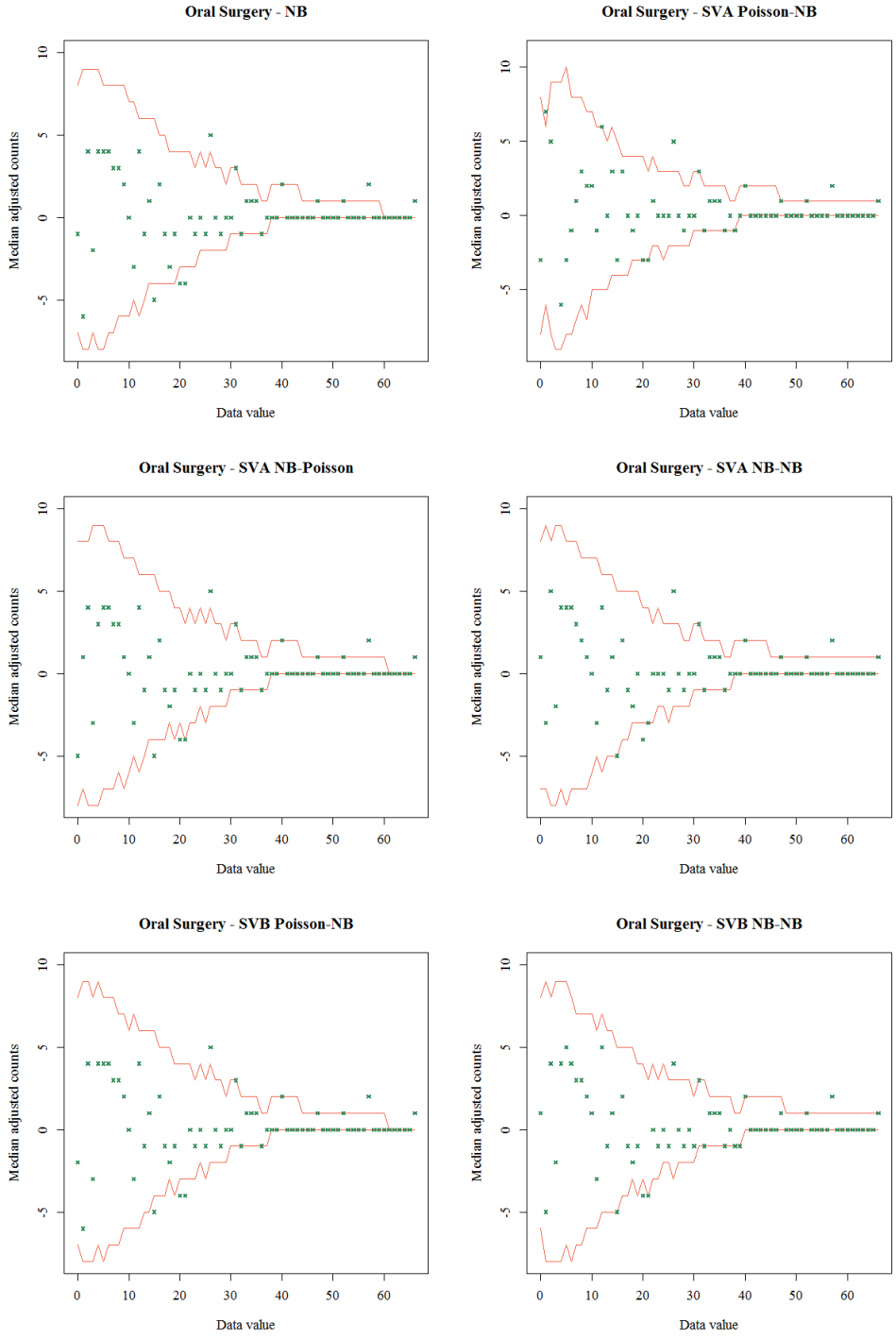

Figure H.19: *Christmas tree plots for Oral Surgery.*

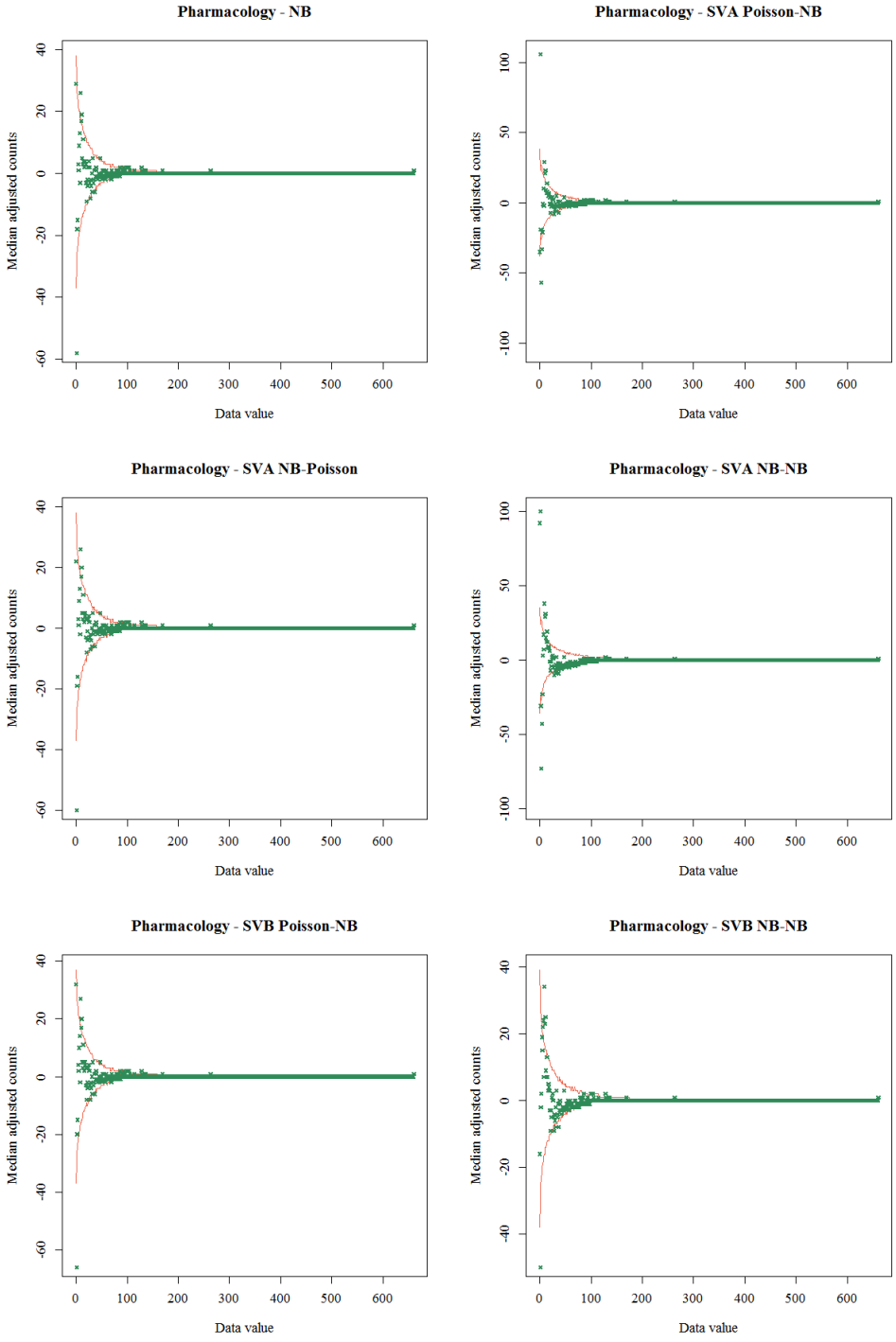

Figure H.20: *Christmas tree plots for Pharmacology.*

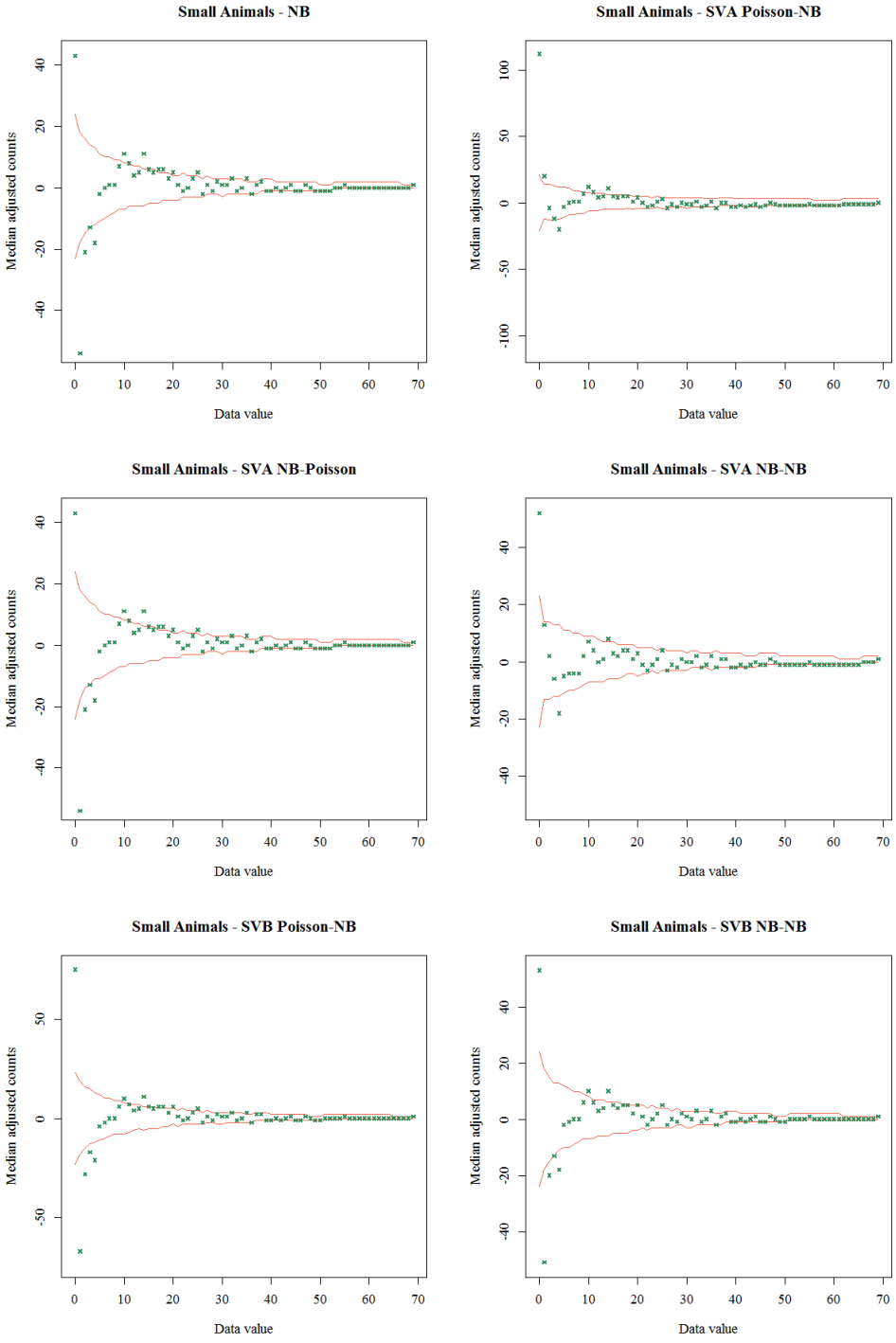

Figure H.21: *Christmas tree plots for Small Animals.*

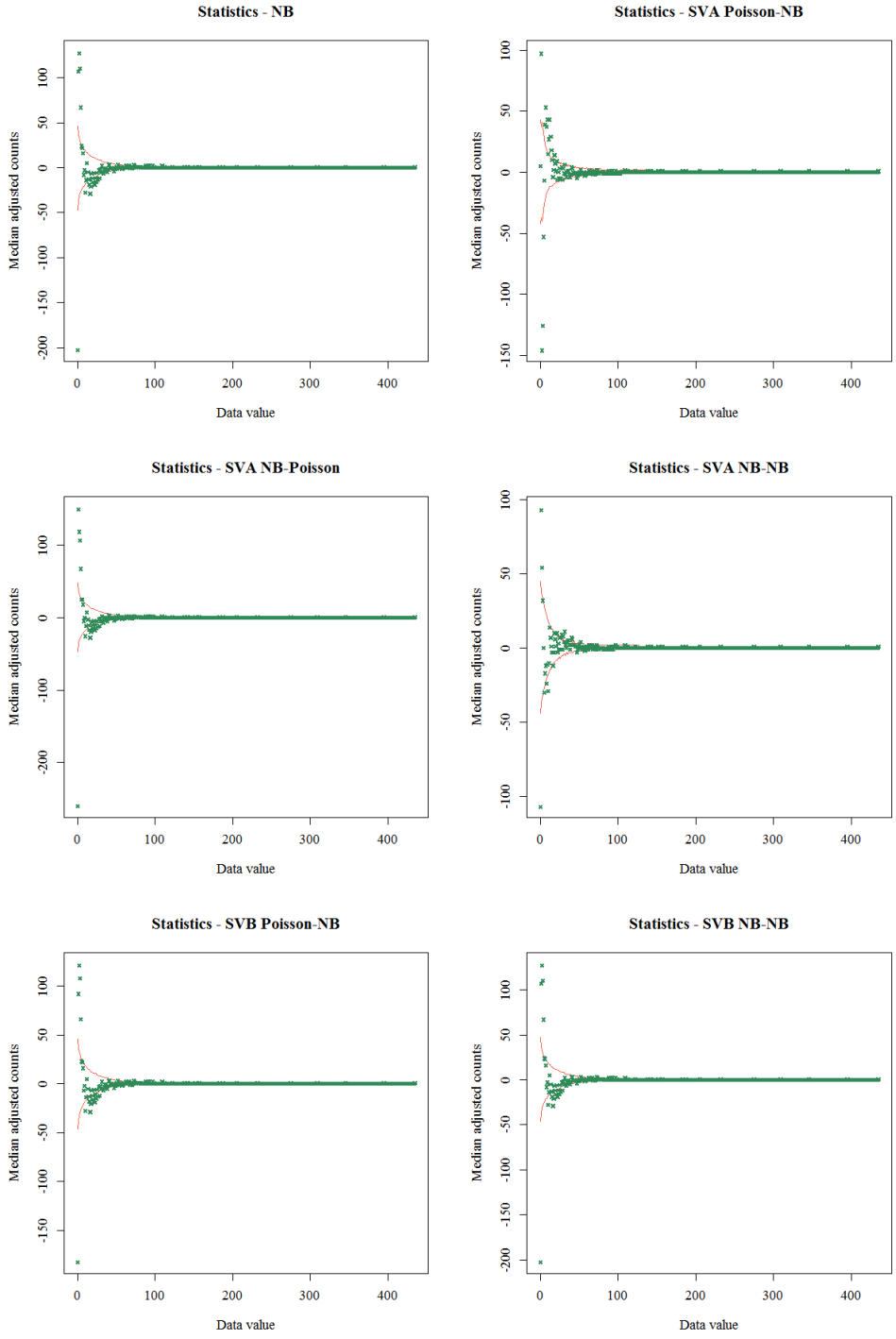

Figure H.22: *Christmas tree plots for Statistics Probability and Uncertainty.*